Simple and Scalable Algorithms for Cluster-Aware Precision Medicine

Amanda M. Buch
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine,
Cornell University
amb2022@med.cornell.edu

Conor Liston
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine,
Cornell University
col2004@med.cornell.edu

Logan Grosenick
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine
Cornell University
log4002@med.cornell.edu

Abstract

AI-enabled precision medicine promises a transformational improvement in healthcare outcomes. However, training on biomedical data presents significant challenges as they are often high dimensional, clustered, and of limited sample size. To overcome these challenges, we propose a simple and scalable approach for cluster-aware embedding that combines latent factor methods with a convex clustering penalty in a modular way. Our novel approach overcomes the complexity and limitations of current joint embedding and clustering methods and enables hierarchically clustered principal component analysis (PCA), locally linear embedding (LLE), and canonical correlation analysis (CCA). Through numerical experiments and real-world examples, we demonstrate that our approach outperforms fourteen clustering methods on highly underdetermined problems (e.g., with limited sample size) as well as on large sample datasets. Importantly, our approach does not require the user to choose the desired number of clusters, yields improved model selection if they do, and yields interpretable hierarchically clustered embedding dendrograms. Thus, our approach improves significantly on existing methods for identifying patient subgroups in multiomics and neuroimaging data and enables scalable and interpretable biomarkers for precision medicine.

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

1 INTRODUCTION

In modern medicine, interpretable clustering of patients into distinct subtypes is increasingly important for personalized biomarker discovery, diagnosis, prognosis, and treatment selection (Sørlie et al., 2001, Santos et al., 2015, Drysdale et al., 2017, Singh and Pandey, 2018, Qian et al., 2019, Bonacchi et al., 2020, Bishop et al., 2022, Buch et al., 2023). To facilitate adoption by healthcare professionals, we need explainable models for clustering that can be trained even when only limited data is available. However, due to the "curse of dimensionality", similarity metrics (and thus clustering algorithm outcomes) degrade in high dimensions (the "$p > N$" setting common in medical imaging, genomics, and multiomics, where we have p correlated variables and N observations fewer than p). 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 the resulting latent space. The mapping to the low-rank space (e.g., 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 (Drysdale et al., 2017, Danda, 2021, Gharavi et al., 2021, Ciortan and Defrance, 2022).

Unfortunately, such two-stage procedures can lead to suboptimal and hard-to-explain results (Chang, 1983), as in the first stage the embedding may ignore important structure in the data relevant to separating clusters in the second stage (see Fig. 1). These issues motivate a need for joint clustering and embedding methods. Identical concerns extend to multiple datasets (multiview learning problems), where clustering and embedding has also typically been approached as a two-stage process, e.g., canonical correlation analysis (CCA) followed by clustering (Chen et al., 2008, Chen and Schizas, 2013, Drysdale et al., 2017, Du et al., 2017, Ouyang, 2019, Buch et al., 2023). Recently, exciting new methods have emerged for jointly clustering and embedding

data, including cluster-aware feature selection (Wang and Allen, 2021), CCA mixture models (Fern et al., 2005, Lei et al., 2017), non-negative matrix factorization (NMF)-based models (Fogel et al., 2016, Wu and Ma, 2020, Zhou et al., 2021), and a number of neural networks (e.g., Huang et al. (2014), Wang et al. (2015), Yang et al. (2016), Mautz et al. (2020), Shin et al. (2020), Lakkis et al. (2021), Boubekki et al. (2021)). Although pioneering, these existing approaches involve complicated many-objective or deep neural network formulations that prioritize clustering over explainability and tend to perform poorly in limited data cases, so far limiting adoption.

Here, to develop an explainable and scalable approach to joint clustering and embedding relevant to precision medicine, we show a straightforward addition of a convex clustering penalty to standard embedding methods yields a simple, theoretically tractable, and modular approach to joint clustering and embedding that is highly competitive in practice and enjoys theoretical benefits over convex clustering in the "large dimensional limit" (LDL) regime appropriate for $p > N$ data (in the LDL regime $p/N \to c$ for constant c as $p, N \to \infty$) (Johnstone, 2001, Paul, 2007, Benaych-Georges and Nadakuditi, 2011, Dobriban, 2017, Aparicio et al., 2020, Bao and Wang, 2022, Couillet and Liao, 2022).

Main contributions and significance:

1. We introduce a modular cluster-aware embedding strategy appropriate for precision medicine applications along with 3 fast/scalable algorithms that solve linear, locally linear, and multiview instantiations.
2. We prove PCMF dominates convex clustering in the LDL regime appropriate for high dimensional data.
3. Our approach does not require specifying cluster number. Instead it outputs *interpretable* per-cluster embeddings organized as a dendrogram.
4. Still, we introduce a model selection procedure for our approach that dominates standard methods.
5. Our approach performs competitively against state-of-the-art methods on 17 real-world datasets.

2 RELATED WORK: CONVEX CLUSTERING

Classically, solving clustering problems using discrete optimization is known to be NP-hard. However, by relaxing the hard clustering constraint to a convex penalty (Pelckmans et al., 2005), 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 penalty, λ. The number of clusters is indirectly controlled by this hyperparameter, and when solved along a path of λ values, convex clustering converges (Radchenko and Mukherjee, 2017, Chi and Steinerberger, 2019) and can exactly recover true data partitions among a mixture of Gaussians (Hocking et al., 2011, Lindsten et al., 2011, Jiang et al., 2020). Further, the solution path can be visualized as a dendrogram to reveal hierarchical structure among clusters (Weylandt et al., 2020).

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 finds estimate $\widehat{X}$ by solving:

$$\underset{\widehat{X} \in \mathbb{R}^{N \times p}}{\text{minimize}} \ \frac{1}{2}\|X - \widehat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q. \quad (1)$$

Tuning λ thus trades off a data approximation term with a convex clustering penalty (which comes from a Lagrangian relaxation of an inequality constraint on the sum of the convex q-norms of differences between the approximated observations—typically $q \in \{1, 2, \infty\}$). Importantly, weights $w_{ij} > 0$ constrained to be nonzero for nearest neighbors (Chi and Lange, 2015, Wang and Allen, 2021) can speed up optimization and increase flexibility in modeling local structure in the row differences, such as with a radial basis function ($w_{ij} = \exp(-\gamma \|X_{i\cdot} - X_{j\cdot}\|_2^2)$) (Hocking et al., 2011, Chi and Lange, 2015), multiplicative weights (Jiang et al., 2020), or properly scaling kernels (Fodor et al., 2022).

Recent theoretical and algorithmic developments for convex clustering (Tan and Witten, 2015, Chiquet et al., 2017, Panahi et al., 2017, Sui et al., 2018, Weylandt, 2019, Jiang et al., 2020, Lin and Chen, 2021, Sun et al., 2021, Fodor et al., 2022) improve practically and theoretically on solving the problem Eq. (1) (Hocking et al., 2011, Chi and Lange, 2015, Panahi et al., 2017, Weylandt et al., 2020, Sun et al., 2021). Crucially, a warm-started ADMM (alternating direction method of multipliers (Glowinski and Marroco, 1975, Gabay and Mercier, 1976, Boyd et al., 2011) approach—Algorithmic Regularization—was recently introduced to enable feasible computation of dense convex clustering λ paths, speeding convergence more than 100-fold (Weylandt et al., 2020). Multiple studies have extended convex clustering, leading to new approaches to biclustering (Allen et al., 2014, Chi and Lange, 2015), multiview clustering (Wang and Allen, 2021), and supervised convex clustering (Wang et al., 2023). These existing approaches do not allow the same variables to contribute differently to multiple clusters, and do not use the convex clustering penalty for joint clustering and embedding, as we do here.

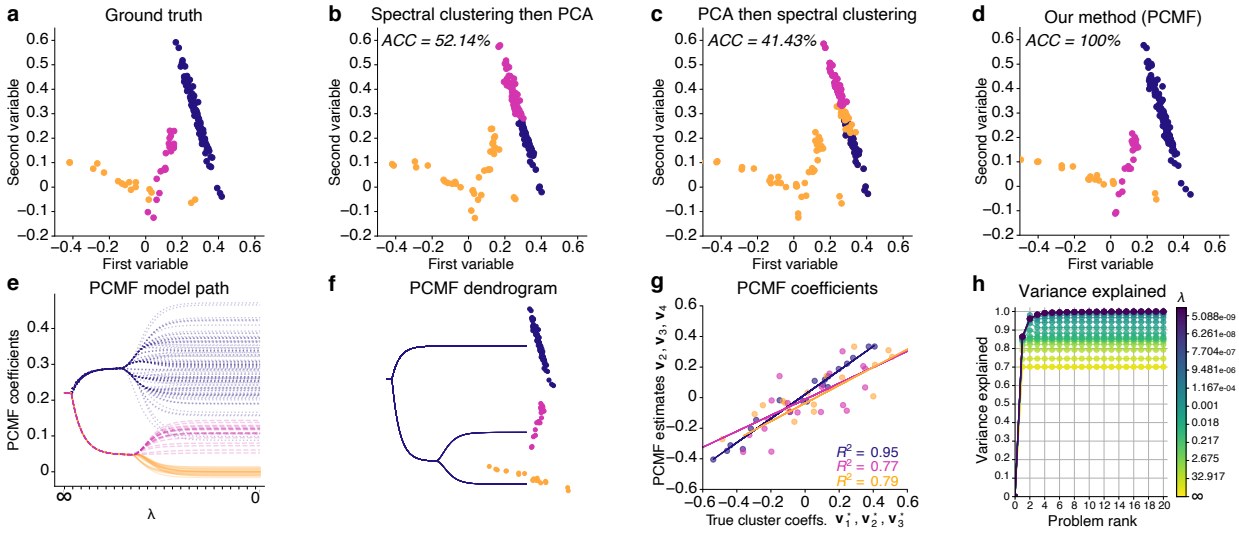

Figure 1: PCMF for explainable joint PCA and hierarchical clustering. **a.** Scatterplot of reconstructed ground truth data (scatterplots show PCA reconstruction of rank $r = 4$) for 3-class problem; $p = 20$; $N_1 = 100$ (blue), $N_2 = 25$ (red), $N_3 = 25$ (orange), colored by true cluster membership. **b.** Spectral clustering sequentially followed by PCA ($r = 4$) on clustered data. **c.** PCA ($r = 4$) sequentially followed by spectral clustering on PCA components. **d.** 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. **b-d.** Color indicates predicted clusters. **e.** PCMF paths for variable 1 fit along decreasing penalty path ($\lambda = \infty$ to $\lambda = 0$). **f.** Interpretable PCMF dendrogram estimated from paths. **g.** 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), $\mathbf{v_2^*}$ (red), and $\mathbf{v_3^*}$ (orange); $\mathbf{v_1}$ corresponds to cluster means intercept vector (not shown). **h.** Calculating variance explained by each PCMF component shows the rank $r = 4$ model correctly captures the three cluster slopes and 1D-direction along which cluster means vary. ACC, Accuracy; Coeffs., coefficients; PCA, principal component analysis

3 OUR APPROACH: PCMF

3.1 Pathwise Clustered Matrix Factorization (PCMF) problem formulation

We propose using the convex clustering penalty as a modular addition to common embedding methods to make them cluster-aware (that is, to enable them to jointly cluster and embed). More explicitly, given a data matrix $X \in \mathbb{R}^{N \times p}$ (with N observations in the rows, p variables in the columns, and rank $R \leq \min(N, p)$), an embedding of X: $\widehat{X} \in \mathcal{M}$ (where $\mathcal{M}$ is a low-dimensional manifold), and a loss function $\mathcal{L}(\cdot, \cdot) : \mathbb{R}^{N \times p} \times \mathbb{R}^{N \times p} \to \mathbb{R}_+$, we can express our general problem as:

$$\underset{\widehat{X} \in \mathcal{M}}{\text{minimize}} \; \mathcal{L}(X, \widehat{X}) + \lambda \sum_{i<j} w_{ij} \|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q, \tag{2}$$

where the penalty term is identical to that used in convex clustering above but now applied to a jointly embedded $\widehat{X} \in \mathcal{M}$. To demonstrate the utility of this strategy, we let $\mathcal{M}$ be the set of rank-r matrices, $\mathcal{M}_r$, and begin with among the most well-known and widely-employed embedding algorithms: the truncated singular value decomposition (tSVD) (Eckart and Young, 1936). In this case, expressing the embedding constraint $\widehat{X} \in \mathcal{M}_r$ explicitly in terms of the tSVD, Eq. (2) yields the PCMF problem:

$$\begin{aligned} &\underset{\widehat{X}, U_r, S_r, V_r}{\text{minimize}} \; \frac{1}{2}\|X - \widehat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q \\ &\text{subject to} \;\; \widehat{X} - U_r S_r V_r^T = 0, \; U_r^T U_r = V_r^T V_r = I_r, \\ &S_r = \text{diag}(s_1, \ldots, s_r), \end{aligned} \tag{3}$$

for $s_1 \geq s_2 \geq \cdots \geq s_r > 0$. Here the rank-$r \leq R$ tSVD embedding is given by $\widehat{X} = U_r S_r V_r^T$, subject to the usual orthogonality constraints on the first r left and right singular vectors (collected in U_r and V_r, respectively) and the standard ordering of the first r singular values on the diagonal of S_r (Eckart and Young, 1936). Without loss of generality, we assume X has been centered—a case where the tSVD is also called principal components analysis (PCA). Note that when $r = R$ (that is, if $\text{rank}(\widehat{X}) = \text{rank}(X)$), this problem reduces to the standard convex clustering problem Eq. (1) as a special case. We next present efficient algorithmic approaches to solving this nonconvex problem.

3.2 PCMF dendrograms for explainability and model selection

PCMF fits a path of solutions along a sequence of values of λ (Fig. 1e–f), and when using the ℓ_2-norm ($q = 2$) (as we do below, given its desirable rotational symmetry), not all members of a cluster are shrunk to exactly the same value (Hocking et al., 2011). Previous work has forced hard clustering at each agglomerative stage along the λ path (Hocking et al., 2011, Weylandt et al., 2020, Jiang et al., 2020). This may artificially force observations into one cluster that may then later switch to another, resulting in nonsmooth paths in practice. We instead introduce letting the paths be unconstrained and smooth while solving divisively, and then to 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 λ is performed on the weighted affinity matrix generated from differences matrix defined by the dual variables as recommended in Chi and Lange (2015). Thus, this procedure estimates the connected components of the affinity graph defined by the dual variables at each value of λ. Further details on model selection are described in Appendix §2.9.

3.3 Solving PCMF with Algorithmic Regularization

Because in most cases it is desirable for many weights w_{ij} in the convex clustering penalty to be exactly zero (Chi and Lange, 2015), we first re-represent the relevant nonzero distances more efficiently as a sparse graph, G. We then introduce an auxiliary variable $G = D\widehat{X} \in \mathbb{R}^{|\mathcal{E}|\times p}$, where $D \in \mathbb{R}^{|\mathcal{E}|\times N}$ is a sparse matrix containing the weighted pairwise distances defined by edges $\mathcal{E}$. This allows us to rewrite the PCMF problem as:

$$\begin{aligned} &\underset{\widehat{X},G,U_r,S_r,V_r}{\text{minimize}} \ \frac{1}{2}\|X-\widehat{X}\|_F^2 + \lambda\sum_{\ell\in\mathcal{E}} w_\ell \|G_{\ell\cdot}\|_q \\ &\text{subject to } \widehat{X} - U_r S_r V_r^T = 0,\ G - D\widehat{X} = 0, \\ &U^TU = V^TV = I_r,\ S_r = \text{diag}(s_1,\ldots,s_r), \end{aligned} \tag{4}$$

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 resulting updates. Critically, we have added Algorithmic Regularization (Weylandt et al., 2020) along the λ path. ADMM solutions fit along a path of λs benefit from "warm-starting" by initializing the next problem along the path at the previous solution. Algorithmic Regularization (AR) takes this to the extreme, shortening steps along the path and decreasing the number of ADMM iterations at each point to a small number (making K small in Algorithm 1). For an appropriately chosen step size, this has been proven to converge to the true path solutions and to speed up the computation of path estimation by $>$ 100-fold (Weylandt et al., 2020). This significantly improves computational feasibility as our algorithm requires solving over many path penalty (λ) values (see Appendix §2 for derivation, convergence details, computational complexity, and consensus algorithm).

Algorithm 1 PCMF

Input: data X, path $\{\lambda\}$, weights $\mathbf{w}$, $\rho \geq 1$,
Notation: data mean $\overline{X}$, rank r, iteration k,
norm $q \in \{1, 2, \infty\}$, pairwise distance matrix D,
proximal operator of $P_{\mathbf{w},q}(\cdot)$: $\text{prox}_{\frac{\lambda}{\rho}P_{\mathbf{w},q}(\cdot)}$

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

3.4 A nonlinear extension: locally linear PCMF (LL-PCMF)

Next, we introduce a locally linear PCMF problem and a Penalized Alternating Least Squares (PALS) (Roweis and Saul, 2000) algorithm to solve it. For clarity (and without loss of generality), we center and scale X, set $s_1 = 1$, and consider the rank-1 version of the PCMF problem (which can be generalized to rank-r using an appropriate deflation approach; see Appendix §2.3). Then denoting the ith column vector of X^T as $\mathbf{x}_i = (X^T)_{\cdot i}$ and defining penalty $\tilde{P}_{\mathbf{w},q}(\mathbf{u},\mathbf{v}) = \sum_{(i,j)\in\mathcal{E}} w_{ij}\|u_i\mathbf{v} - u_j\mathbf{v}\|_q$, we can write the rank-1 tSVD with a convex clustering penalty (see Appendix §2.3) as:

$$\begin{aligned} &\underset{\mathbf{u},\mathbf{v}}{\text{minimize}} \ \sum_{i=1}^N \frac{1}{2}\|\mathbf{x}_i - u_i\mathbf{v}\|_2^2 + \lambda\tilde{P}_{\mathbf{w},q}(\mathbf{u},\mathbf{v}) \\ &\text{subject to } \|\mathbf{u}\|_2^2 = 1,\ \|\mathbf{v}\|_2^2 = 1. \end{aligned} \tag{5}$$

To introduce the overparameterization necessary for convex clustering we replace the single vector $\mathbf{v}$ with a matrix $V \in \mathbb{R}^{p\times N}$ containing column vectors $\mathbf{v}_i = V_{\cdot i}$ (denoting the set of these column vectors as $\{\mathbf{v}\}_i,\ i = 1,\dots,N$)—this allows each observation to potentially be its own cluster in the limit $\lambda \to 0$. Note this is the same overparameterization as in the standard convex clustering problem Eq. (1). 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 the overparameterized problem:

$$\begin{aligned} &\underset{\mathbf{u},V}{\text{minimize}} \sum_{i=1}^{N} \frac{1}{2}\|\mathbf{x}_i - u_i\mathbf{v}_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}, V) \\ &\text{subject to } \|\mathbf{u}\|_2^2 = 1,\ \|\mathbf{v}_i\|_2^2 = 1,\ i = 1,\dots,N. \end{aligned} \tag{6}$$

Next, by removing the cross-terms in the penalty we allow $\mathbf{u}$ and $\mathbf{v}$ to independently vary and the locally-defined weights w_{ij} to apply to the embedding (making it locally linear). We do this by replacing $P_{\mathbf{w},q}(\mathbf{u}, \mathbf{v})$ with $Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u}) = \sum_{(i,j)\in\mathcal{E}} w_{ij}|u_i - u_j|$ and $Q^{V}_{\mathbf{w},q}(V) = \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_q$. Then using fixed values from iterate k, $y^k_{\mathbf{u},i} = \mathbf{x}_i^T\mathbf{v}_i^k$ and $\mathbf{y}^k_{\mathbf{v},i} = u_i^k\mathbf{x}_i$, we get updates:

$$\begin{aligned} &\mathbf{u}^{k+1} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \sum_{i=1}^{N} \|y^k_{\mathbf{u},i} - u_i\|_2^2 + \lambda Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u}) \\ &\text{subject to } \|\mathbf{u}\|_2^2 = 1, i = 1,\dots,N, \end{aligned} \tag{7a}$$

$$\begin{aligned} &\{\mathbf{v}_i\}^{k+1} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^{N} \|\mathbf{y}^k_{\mathbf{v},i} - \mathbf{v}_i\|_2^2 + \lambda Q^{V}_{\mathbf{w},q}(V) \\ &\text{subject to } \|\mathbf{v}_i\|_2^2 = 1,\ i = 1,\dots,N. \end{aligned} \tag{7b}$$

Note these iterative PALS updates for LL-PCMF are just convex clustering problems with constraints, and thus given some convex clustering solver ConvexCluster (Appendix Algorithms 4–5), we arrive at our algorithm for LL-PCMF (see Appendix Algorithm 3 and Appendix §2.3 for algorithm and derivation).

3.5 A multiview extension: Pathwise Clustered CCA (P3CA)

We next extend our approach to multiview learning, where we aim to jointly learn low-rank correlation structure while clustering observations across multiple data views (i.e., fitting canonical correlation analysis or CCA within clusters). To do so, we follow a derivation similar to LL-PCMF (note it is also straightforward to derive a linear P3CA by instead replacing the SVD with CCA in Alg. 1), introducing the overparameterized pathwise clustered canonical correlation analysis (P3CA) optimization problem (recall $\mathbf{v}_i = V_{\cdot i}$ are column vectors of $V \in \mathbb{R}^{p\times N}$). We have data matrices $X \in \mathbb{R}^{N\times p_X}, Y \in \mathbb{R}^{N\times p_Y}$, and variables $\mathbf{u}_i \in \mathbb{R}^{p_X}$, $\mathbf{v}_i \in \mathbb{R}^{p_Y}$, and we define $\Sigma_i = X_{i\cdot}^T Y_{i\cdot} \in \mathbb{R}^{p_X\times p_Y}$ and $Q_{\mathbf{w},q}(V) = \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_q$. This yields the penalized rank-1 CCA problem:

$$\begin{aligned} &\underset{\{\mathbf{u}_i\},\{\mathbf{v}_i\}}{\text{minimize}} -\sum_{i=1}^{N} \mathbf{u}_i^T\Sigma_i\mathbf{v}_i\ + \lambda Q_{\mathbf{w},q}(U) + \lambda Q_{\mathbf{w},q}(V) \\ &\text{subject to } \|\mathbf{u}_i\|_2^2 = 1,\ \|\mathbf{v}_i\|_2^2 = 1, \end{aligned} \tag{8}$$

for $i = 1,\dots,N$. Without inequality constraints, this is a biconvex problem in the $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ when the subproblems are 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:

$$\begin{aligned} &\{\mathbf{u}_i\}^{k+1} \leftarrow \underset{\{\mathbf{u}_i\}}{\text{argmin}} \sum_{i=1}^{N} \frac{1}{2}\|\tilde{\mathbf{x}}_i - \mathbf{u}_i\|_2^2 + \lambda Q_{\mathbf{w},q}(U) \\ &\text{subject to } \|\mathbf{u}_i\|_2^2 = 1,\ i = 1,\dots,N, \end{aligned} \tag{9a}$$

$$\begin{aligned} &\{\mathbf{v}_i\}^{k+1}\ \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^{N} \frac{1}{2}\|\tilde{\mathbf{y}}_i - \mathbf{v}_i\|_2^2 + \lambda Q_{\mathbf{w},q}(V) \\ &\text{subject to } \|\mathbf{v}_i\|_2^2 = 1,\ i = 1,\dots,N. \end{aligned} \tag{9b}$$

Each update is again a constrained convex clustering problem, leading to Algorithm 2. Empirically, for sufficiently small steps sizes, Algorithmic Regularization closely approaches the ADMM solutions with a significant speed up (see Appendix §2.1 and §2.8 for derivation and computational complexity).

Algorithm 2 Pathwise Clustered Canonical Correlation Analysis (P3CA)

Input: data (X, Y), path $\{\lambda\}$, weights $\mathbf{w}$, $\rho \geq 1$,
Notation: iter. k, data means $(\bar{X}, \bar{Y})$, $\mathbf{v}_i = V_{\cdot i}$, $\tilde{\mathbf{x}}_i = (\tilde{X}_{i\cdot})^T$, $\tilde{\mathbf{y}}_i = (\tilde{Y}_{i\cdot})^T$, norm $q \in \{1, 2, \infty\}$
$U \leftarrow \bar{X}, V \leftarrow \bar{Y}$
for $\lambda \in \{\lambda\}$ **do**
  for $k = 1,\dots,K$ **do**
    $\tilde{\mathbf{x}}_i^{k+1} \leftarrow \Sigma_i\mathbf{v}_i^k$ $(\Sigma_i = X_{i\cdot}Y_{i\cdot}^T \in \mathbb{R}^{p_X\times p_Y})$ for $i = 1,\dots,N$
    $\mathbf{u}_i^{k+\frac{1}{2}} \leftarrow$ ConvexCluster$(\tilde{X}^{k+1}, U^k, \lambda, \mathbf{w}, \text{q})$
    $\mathbf{u}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}_i^{k+\frac{1}{2}})$ for $i = 1,\dots,N$
    $\tilde{\mathbf{y}}_i^{k+1} \leftarrow \Sigma_i^T\mathbf{u}_i^{k+1}$ $(\Sigma_i^T = Y_{i\cdot}X_{i\cdot}^T \in \mathbb{R}^{p_Y\times p_X})$ for $i = 1,\dots,N$
    $\mathbf{v}_i^{k+\frac{1}{2}} \leftarrow$ ConvexCluster$(\tilde{Y}^{k+1}, V^k, \lambda, \mathbf{w}, \text{q})$
    $\mathbf{v}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}})$ for $i = 1,\dots,N$
  end for
  Save path solutions: $U_{i\cdot}^K \leftarrow \mathbf{u}_i^{KT}$; $V_{i\cdot}^K \leftarrow \mathbf{v}_i^{KT}$ for $i = 1,\dots,N$; $(U_\lambda, V_\lambda) \leftarrow (U^K, V^K)$
  Initialize: $(U^0, V^0) \leftarrow (U^K, V^K)$
end for
return pathwise solutions $\{U_\lambda\}, \{V_\lambda\}$

3.6 Theory: PCMF for Gaussian Mixture Model (GMM) data in the Large Dimensional Limit (LDL) regime

Here we show that our approach dominates convex clustering in the $p > N$ LDL regime relevant to precision medicine, offering a constructive proof for inadmissibility of convex clustering in the case of "nontrivially" clustered GMM data following recent results from random matrix theory (RMT) (Couillet and Liao, 2022).

Definition 1 "Nontrivial" GMM data. We observe N i.i.d. data vectors $\mathbf{x}_i \in \mathbb{R}^p$ drawn from the K-class GMM with fixed class sizes $N_1, \ldots, N_K$ (with $\sum_{k=1}^K N_k = N$) gathered in data matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times p}$, with $p \sim N$ or $p > N$ such that $p/N \to c \in (0, \infty)$ and $N_a/N \to c_a \in (0, 1)$ as $N, N_a, p \to \infty$. Letting $\mathcal{C}_a$ be the set of observations from class a for $a \in \{1, \ldots, K\}$ such that $\mathbf{x}_i \sim \mathcal{N}(\mu_a, C_a) \iff \mathbf{x}_i \in \mathcal{C}_a$ with $C_1, \ldots, C_k$ distinct and of bounded norm. To ensure that cluster separation is nontrivial as $N, p \to \infty$ we take $\|\mathbf{x}_i\|$ to be of order $O(\sqrt{p})$ and $\|\mu_a - \mu_b\| = O(1)$ for $a, b \in 1, \ldots, K; a \neq b$. See Appendix §3 for additional details and motivation.

Proposition 1 For clustering the nontrivial GMM data in the LDL regime, PCMF asymptotically dominates standard convex clustering.

Proposition 2 For clustering the nontrivial GMM data in the LDL regime, the local linear relaxation LL-PCMF asymptotically dominates convex clustering.

Proposition 3 PCMF generalizes kernel spectral clustering (kSC) to joint clustering and embedding.

We briefly explain these results here leaving proofs to the Appendix §3. To prove Proposition 1, we let $C^\circ = \sum_{a=1}^k \frac{N_a}{N} C_a$ and note that due to "universality" results from RMT (see Couillet and Liao (2022) Ch. 2) the GMM assumption is often (though not always) equivalent to requiring $x_i \in \mathcal{C}_a : x_i = \mu_a + C_a^{1/2} z_i$ (where z_i is a random vector with i.i.d. zero mean, unit variance, and suitably bounded higher-order moment entries) (Couillet and Liao, 2022). We then consider the convex clustering penalty "element-wise" and find that:

$$\sum_{i,j \in \mathcal{E}} f\left(\frac{1}{p}\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\|_2^2\right) = \sum_{i,j \in \mathcal{E}} f\left(\frac{2}{p} C^\circ + O(p^{-1/2})\right), \tag{10}$$

where we have subsumed weights w_{ij} and the square root into $f(\cdot)$ and normalized by p. The equality in Eq. (10) follows from expanding and considering each term given these assumptions, and indicates if we consider entry-wise distances in the LDL regime, all entries are dominated by the constant $O(1)$ term $2\text{tr}\, C^\circ/p$ regardless of the values of a and b. Further $\|\mu_a - \mu_b\|_2^2/p = O(p^{-1})$ is dominated by the $O(p^{1/2})$ noise terms, indicating convex clustering's entry-wise distance approach does not allow discrimination here.

However, by instead considering the data "matrix-wise" via embedding, we find that there is important discriminative information available in the low-rank structure of the Euclidean distances. In particular, although the matrix is again dominated by a $O(N)$-norm constant matrix $2\text{tr}\, C^\circ/p \cdot \mathbf{1}_N \mathbf{1}_N^T$, this rank-1 term can be discarded leaving resolvable spectral information about the covariance traces in the $O(N^{-1/2})$-norm rank-2 second dominant term. Subsequent order $O(1)$ terms contain usable discriminative information about the means. Thus, using a low-rank embedding in the convex clustering term (PCMF) allows discrimination in the nontrivial GMM case and standard convex clustering does not. The proof of Proposition 2 follows similarly. Proposition 3 follows from noting that the differences in the penalty term can be represented through an application of the Laplacian of the weight-induced graph such that PCMF is jointly optimizing both a (kernel) spectral embedding and a clustering on that embedding (see Appendix §3 for proofs). Together, these propositions show PCMF outperforms convex clustering in $p > N$ problems and formalize its relationship to kSC.

4 EXPERIMENTAL SETUP

We measure clustering accuracy (ACC) using PCMF, LL-PCMF, and P3CA against 14 other clustering methods on 17 real-world datasets (Table 1): (1) PCA/CCA + K-means (Hotelling, 1933, 1936, MacQueen, 1967, Hastie et al., 2009), (2) Ward (Hastie et al., 2009), (3) Spectral (Hastie et al., 2009), (4) Elastic Subspace (You et al., 2016a,b), (5) gMADD (Sarkar and Ghosh, 2020, Paul et al., 2021), (6) HDCC (Bouveyron et al., 2007, Bergé et al., 2012), (7) Leiden (Traag et al., 2019), (8) Louvain (Blondel et al., 2008), (9) DP-GMM (Escobar and West, 1995, You et al., 2016a), (10) convex clustering (hCARP) (Weylandt et al., 2020), (11) PCA/CCA + convex clustering hCARP, (12) Deep Embedding Clustering (DEC) (Xie et al., 2016), (13) IDEC (Guo et al., 2017), and (14) CarDEC (Lakkis et al., 2021). See Appendix §5 for hyperparameters and additional experimental details.

5 EXPERIMENTAL RESULTS

We assessed the efficacy of our unsupervised cluster-aware embedding method on a total of 39 datasets comprised of 22 synthetic datasets and 17 biomedical datasets. We measured the clustering accuracy (ACC; accuracy of clustering vs. ground-truth clusters) of PCMF, LL-PCMF, and P3CA in comparison to 15

Table 1: 17 real-world datasets in Main Text Tables 2-3 and cluster discovery analysis.

Dataset	Variables (p)	Samples (N)	Classes
(1) NCI	6,830 genes (expression) (X)	64	13 cell types
(2) SRBCT	2,318 genes (expression) (X)	88	4 cancer diagnoses
(3) Mouse	16,944 genes (scRNA-seq) (X)	125	7 mouse organ types
(4) Tumors	11,931 expression/methylation (X)	142	3 cancer diagnoses
(5) Tumors-Large	11,931 expression/methylation (X)	400	3 cancer diagnoses
(6) Monkey-LGN	45,768 genes (expression) (X)	1,801	2 cell types
(7) Mouse-LGN	39,670 genes (expression) (X)	1,818	2 cell types
(8) MNIST	784 image pixels from 28 x 28 pixel image (X)	36,000	6 digit types
(9) MNIST Fashion	784 image pixels from 28 x 28 pixel image (X)	36,000	6 clothing types
(10) Human-ATAC	21,972 chromatin profiles (X)	30,480	2 cell types
(11) COVID-19 (Multiview)	403 metabolites (X); 382 proteins (Y)	45	3 severities
(12) NCI (Multiview)	1,000 genes (expression) (X); 100 genes (expression) (Y)	64	13 cell types
(13) SRBCT (Multiview)	1,000 genes (expression) (X); 100 genes (expression) (Y)	88	4 cancer diagnoses
(14) Mouse (Multiview)	1,000 genes (scRNA-seq) (X); 100 genes (scRNA-seq) (Y)	125	7 mouse organ types
(15) Tumors (Multiview)	1,000 genes (expression) (X); 100 genes (expression) (Y)	142	3 cancer diagnoses
(16) Autism Spectrum Disorder (ASD) (Multiview)	3 behaviors (X); 20 RSFC features (Y)	299	Unknown (discovery analysis)
(17) Palmer Penguin (Multiview)	2 features (X); 2 features (Y)	342	3 penguin species

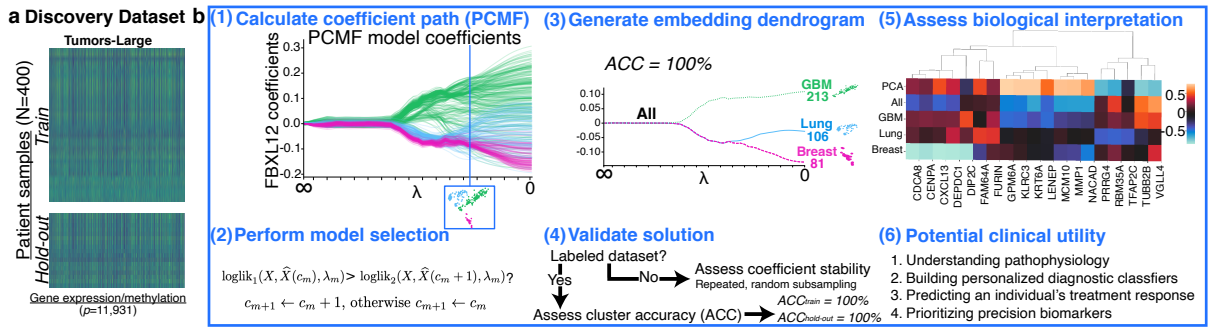

Figure 2: Flow diagram of PCMF on Tumors-Large dataset. **a.** Discovery dataset split into in-sample for training and hold-out for testing. **b.** Flow diagram of cluster-aware embedding and interpretation steps: **1.** Calculate coefficient paths using PCMF. **2.** Perform model selection on paths. **3.** Calculate embedding dendrogram from paths and selected model. **4.** Validate solution using cluster accuracy (if labeled data) and/or PCMF coefficient stability. **5.** Biological interpretation (e.g., top-weighted coefficients as biomarkers, gene set enrichment, protein-protein interaction network (PPI), pathway analysis). **6.** Examples of potential clinical utility.

other clustering approaches, including 3 deep embedded clustering approaches (Figs. 1,2,3,4, Tables 2,3, Appendix Tables 1–6; Appendix Figs. 1–9). One can also use our model selection approach on the PCMF solution paths when the number of clusters is unknown, and evaluating on synthetic and biomedical data we significantly outperform standard cluster number selection methods including Silhouette, Calinski-Harabasz, and Davies-Bouldin statistics (see Appendix §2.9).

In small N biomedical datasets (underdetermined, $p > N$), we found LL-PCMF and P3CA outperform all methods except DEC/IDEC on SRBCT (Table 2, Appendix Table 1,3,4). In 12 synthetic datasets, we found that PCMF and LL-PCMF with nearest neighbors $N.N. = 25$ performs competitively in accuracy, especially for $p > N$ ($p = 200$, $p = 2,000$; Appendix Table 1). To evaluate the PCA interpretation of PCMF embeddings, we compared and showed high similarity to tSVD estimates fit within ground-truth clusters (Fig. 1g, Appendix Table 2, and Appendix Fig. 2).

Next, we evaluated PCMF on large (many observations N) datasets, using a consensus formulation for scalability. Standard convex clustering cannot run on datasets of this size (i.e., $N > 1,000$). In large N biomedical and synthetic datasets, we found PCMF outperforms other methods on in-sample and held-out test set data (Table 3, Fig. 3, Appendix Fig. 2, Appendix Table 2).

In the tumors-large dataset ($N = 400$), we found PCMF model coefficients for the F-Box And Leucine Rich Repeat Protein 2 (FBXL2) gene reveal a cluster hierarchy between GBM, lung, and breast cancer while a two-step approach has degraded cluster membership (Fig. 3a-c). Branching structure reflects the suspected role of FBXL2 as a metastatic biomarker of breast-to-lung metastasis (Wang et al., 2019) and suggests a druggable target (Deng et al., 2020). In Fig. 3d, Spearman's correlations between the PCMF score and prolactin receptor (PRLR) gene expression reveal strong slope differences between the 3 cancer tumors. PRLR is a mammary proto-oncogene (Sa-Nguanraksa et al., 2020, Grible et al., 2021), and a prognostic biomarker of GBM progression (higher expression with shorter

Table 2: Clustering accuracy on small real-world datasets ("MV" abbreviates "Multiview").

	NCI	SRBCT	Mouse	Tumors	COVID-19-MV	Penguins-MV	NCI-MV	SRBCT-MV	Mouse-MV	Tumors-MV
Variables (p)	6,830	2,318	16,944	11,931	403; 382	2; 2	1,000; 100	1,000; 100	1,000; 100	1,000; 100
Samples (N)	64	88	125	142	45	342	64	88	125	142
Classes	13	4	7	3	3	3	13	4	7	3
PCMF	43.79%	51.8%	73.6%	92.25%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
LL-PCMF	**64.06%**	55.42%	**80.00%**	**97.89%**	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
P3CA	n.a.	n.a.	n.a.	n.a.	**91.11%**	**98.25%**	**56.25%**	**65.06%**	**63.20%**	**98.59%**
PCA + K-means	39.06%	40.96%	45.60%	50.00%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CCA + K-means	n.a.	n.a.	n.a.	n.a.	51.11%	79.82%	31.25%	37.35%	27.20%	50.70%
Ward	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%	72.54%	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%	40.06%	48.43%	46.99%	49.60%	71.13%
Louvain	42.19%	48.19%	76.00%	94.34%	82.22%	65.20%	45.31%	48.19%	60.80%	93.66%
DP-GMM	46.88%	43.37%	54.40%	85.92%	73.33%	68.42%	45.31%	44.58%	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%
PCA + hCARP	37.50%	46.99%	47.20%	46.48%	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
CCA + hCARP	n.a.	n.a.	n.a.	n.a.	46.67%	95.32%	29.69%	46.99%	30.40%	38.73%
DEC	45.31%	**71.08%**	46.40%	94.37%	86.67%	88.89%	54.69%	**65.06%**	33.60%	94.37%
IDEC	48.44%	67.47%	61.60%	92.96%	73.33%	n.a.	n.a.	n.a.	n.a.	n.a.
CarDEC	51.56%	40.96%	75.20%	90.14%	84.44%	n.a.	n.a.	n.a.	n.a.	n.a.

Table 3: Clustering accuracy (ACC) and time elapsed (TOC) for consensus PCMF on large datasets. ("X" indicates computationally infeasible to run. "T" indicates infeasible due to run time out.)

	Tumors-Large		Monkey-LGN		Mouse-LGN		MNIST		Fashion MNIST		Human-ATAC		Synthetic	
Variables (p)	11,931		45,768		39,670		784		784		21,972		1,000	
Samples (N)	400		1,801		1,818		36,000		36,000		30,480		100,000	
Classes	3		2		2		6		6		2		4	
	In-sample	Hold-out	In-sample	Hold-out	In-sample	Hold-out	In-sample	Hold-out	In-sample	Hold-out	In-sample	Hold-out	In-sample	Hold-out
PCMF	**100.00%**	**100.00%**	**99.10%**	**98.40%**	**96.10%**	**66.99%**	**99.93%**	**88.33%**	**99.94%**	**81.41%**	**84.60%**	**86.51%**	**100.00%**	**100.00%**
PCA + K-means	89.75%	**100.00%**	68.20%	92.81%	66.61%	51.23%	29.64%	29.64%	45.00%	45.48%	78.32%	77.95%	50.09%	50.00%
Ward	90.50%	n.a.	67.80%	n.a.	86.20%	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.
Spectral	92.00%	n.a.	65.50%	n.a.	97.30%	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.
gMADD	61.50%	n.a.	85.10%	n.a.	60.60%	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.	–X–	n.a.
Leiden	66.25%	n.a.	28.70%	n.a.	31.20%	n.a.	60.62%	n.a.	38.31%	n.a.	49.40%	n.a.	10.88%	n.a.
Louvain	72.25%	n.a.	30.10%	n.a.	39.00%	n.a.	69.88%	n.a.	42.26%	n.a.	51.50%	n.a.	10.85%	n.a.
DEC	99.25%	n.a.	93.90%	n.a.	92.50%	n.a.	–T–	n.a.	–T–	n.a.	86.60%	n.a.	–T–	n.a.
IDEC	86.50%	n.a.	–T–	n.a.	–T–	n.a.	55.25%	n.a.	48.98%	n.a.	–T–	n.a.	–T–	n.a.

survival in males) (Asad et al., 2020) and therapeutic target (Asad et al., 2019, Sa-Nguanraksa et al., 2020). PRLR is strongly but oppositely associated with GBM ($R = -0.81$) and breast tumor clusters ($R = 0.45$), as suggested for triple-negative breast cancer–shows higher expression is associated with lower recurrence and longer survival (Motamedi et al., 2020).

Next, in a small N COVID-19 dataset (Shen et al., 2020) we show that P3CA identifies hierarchical clustered metabolome-proteome embeddings that predict both severity ($ACC = 91.11\%$ in Table 2) and potential biomarkers. Severity hierarchy is not well represented in the two-step approach; Fig. 4a-b. Cluster-specific P3CA score Spearman's correlations with Carboxypeptidase B2 (CPB2) and Apolipoprotein M (APOM) proteins (Fig. 4c-d) show opposite relationships—CPB2 (a known predictor of severe illness and a therapeutic target (Foley et al., 2015, Zhang et al., 2021, Claesen et al., 2022) is strongly associated with the severe cluster ($R = -.8$). APOM (known to associate with less severity and better prognosis (Shen et al., 2020, Cosgriff et al., 2022) is only strongly associated with the not severe P3CA cluster ($R = 0.69$). Protein-protein interaction (PPI) networks constructed using the top 25 cluster-associated proteins reveal only a small PPI network for the healthy cluster with just 5/25 COVID-19-related genes versus large, highly-connected PPI networks with 15-18 COVID-19-related genes in the others (Fig. 4e-h; see methods in Appendix §5).

Finally, we show P3CA's utility for discovering autism spectrum disorder (ASD) subgroups, relevant to personalized diagnosis (Drysdale et al., 2017, Grosenick et al., 2019, Buch et al., 2023) (Appendix §4.3). We find strong differences in associations of ASD (Martino, 2014, 2017) subtype embeddings with behavior and brain connectivity (Appendix Fig. 9 and Appendix Tables 5–6), consistent with known ASD subpopulation differences on behaviors with prefrontal cortex to somatosensory cortex, posterior parietal cortex, and middle temporal gyrus (Buch et al., 2023). Subject-level P3CA embedding coefficients are robust to data perturbation (cosine similarity: 0.93 ± 0.05 for U estimates and 0.97 ± 0.03 for V estimates; 10 subsamples). Notably, our approach provides more distinct cluster separation along a single P3CA dimension, and improves interpretability relative to current SOTA methods in neuroimaging for this ASD subtyping problem (Buch et al., 2023).

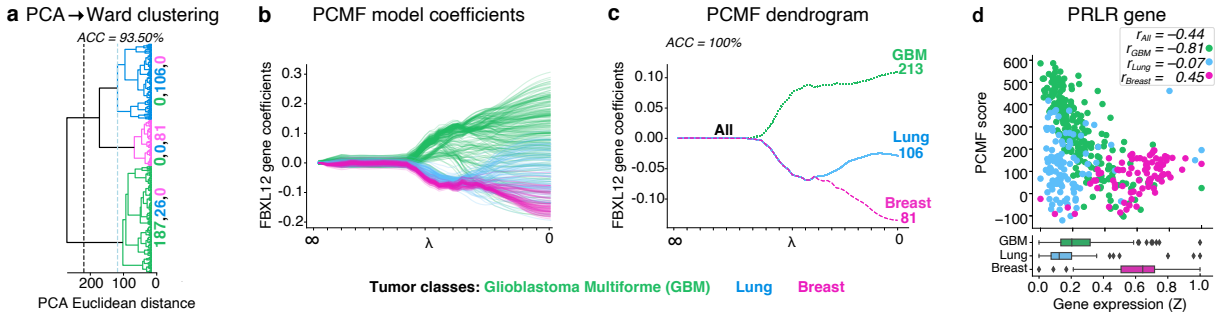

Figure 3: PCMF identifies tumor clusters and embeddings using gene expression ($p = 11,931$) from $N = 400$ samples. **a.** Dendrogram shows hierarchical clustering on PCA embedding (PCA—> Ward clustering). **b.** PCMF path and **c.** dendrogram shows PCMF perfectly recovers clusters ($ACC = 100\%$). **d.** Scatterplots/boxplots show distribution of PRLR gene expression versus PCMF expression scores per sample, colored by PCMF-predicted clusters. r: correlation between gene expression and PCMF expression score. ACC, Accuracy; dist., distance

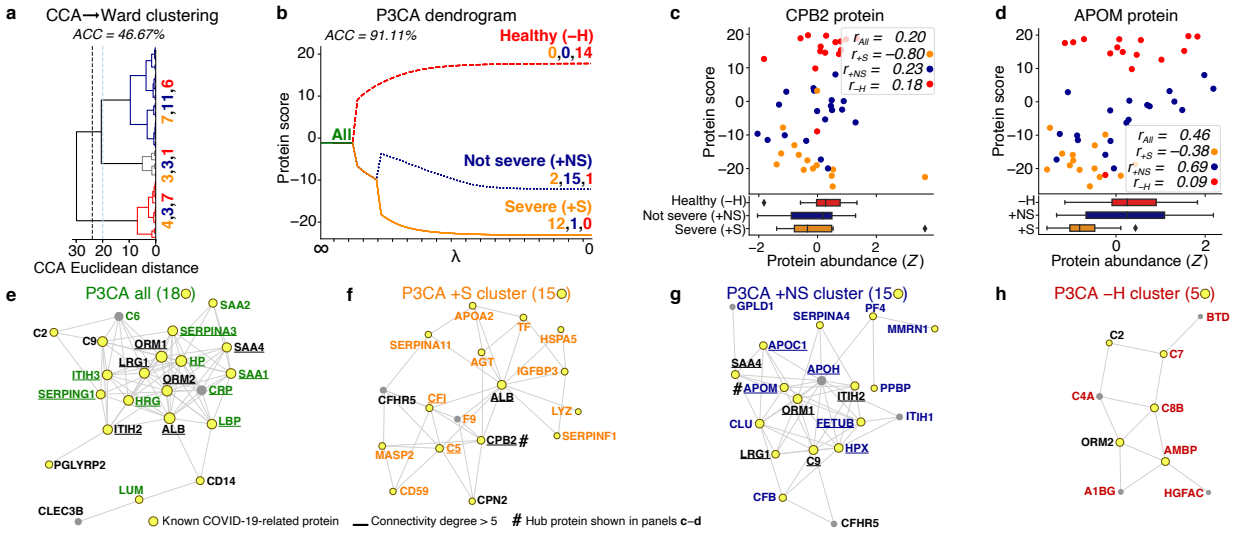

Figure 4: P3CA identifies COVID-19-severity clusters and embeddings using protein ($p = 382$) and metabolite ($p = 403$) abundances from $N = 45$ individuals. **a.** Dendrogram shows hierarchical clustering on CCA embedding (CCA—> Ward clustering) fails to identify clusters. **b.** P3CA dendrogram shows P3CA accurately identifies clusters ($ACC = 91.11\%$). **c-d.** Scatterplots/boxplots show distribution of protein abundance versus P3CA protein scores for each individual colored by P3CA-predicted clusters. r: correlation between abundances and P3CA protein score (i.e., canonical variate). **e-h.** Protein-protein interaction (PPI) networks for top 25 proteins associated with each P3CA-predicted cluster. Colored text: cluster-specific proteins; yellow node: COVID-19-associated; underlined: network degree>5

6 DISCUSSION AND CONCLUSION

AI-enabled precision medicine promises dramatic improvements in healthcare, but facilitating adoption by healthcare professionals will require explainable, effective, and scalable methods appropriate for biomedical data. To meet this need, we have introduced a simple and interpretable joint clustering and embedding strategy using a modular convex clustering penalty. We instantiate our approach in three scalable algorithms that solve linear (PCMF), nonlinear (LL-PCMF), and multiview (P3CA) problems, and show that our method dominates standard convex clustering in the LDL regime. Empirically, our results are highly competitive against classical and SOTA clustering approaches and have superior explainability to SOTA clustering approaches, enabling discovery of an interpretable hierarchy of cluster-wise embeddings that can predict diagnosis- and prognosis-relevant biomarkers. Still they have limitations: they are less flexible than neural network methods and thus may be dominated by such approaches in observation-rich cases. Overall, we present a simple and effective approach to help customize biomarker discovery, diagnosis, prognosis, and treatment selection that is particularly effective in data-limited $p > N$ cases.

Code availability

Code may be found at: https://github.com/CARVE-AI.

Acknowledgements

This work was supported by a Ford Foundation Postdoctoral Fellowship, a New Venture Fund grant (NVF 202423-01), an NIMH Stephen I. Katz Early Stage Investigator Research Project Grant (MH131534), a Cornell Center for Pandemic Prevention and Response Seed Grant, and a Whitehall Foundation Grant.

References

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

L. Aparicio, M. Bordyuh, A. J. Blumberg, and R. Rabadan. A random matrix theory approach to denoise single-cell data. *Patterns (N Y)*, 1(3): 100035, June 2020.

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.

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.

Z. Bao and D. Wang. Eigenvector distribution in the critical regime of BBP transition. *Probab. Theory Related Fields*, 182(1):399–479, Feb. 2022.

F. Benaych-Georges and R. R. Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Adv. Math.*, 2011.

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.

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.

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.

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.

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.

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

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

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.

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.

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.

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.

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

E. C. Chi and S. Steinerberger. Recovering trees with convex clustering. *Siam Journal on Mathematics of Data Science*, 1(3):383–407, 2019.

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

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

K. Claesen, Y. Sim, A. Bracke, M. De Bruyn, E. De Hert, G. Vliegen, A. Hotterbeekx, A. Vujkovic, L. van Petersen, F. H. R. De Winter, I. Brosius, C. Theunissen, S. van Ierssel, M. van Frankenhuijsen, E. Vlieghe, K. Vercauteren, S. Kumar-Singh, I. De Meester, and D. Hendriks. Activation of the

carboxypeptidase U (CPU, TAFIa, CPB2) system in patients with SARS-CoV-2 infection could contribute to COVID-19 hypofibrinolytic state and disease severity prognosis. *J. Clin. Med. Res.*, 11(6), Mar. 2022.

C. V. Cosgriff, T. A. Miano, D. Mathew, A. C. Huang, H. M. Giannini, L. Kuri-Cervantes, M. B. Pampena, C. A. G. Ittner, A. R. Weisman, R. S. Agyekum, T. G. Dunn, O. Oniyide, A. P. Turner, K. D'Andrea, S. Adamski, A. R. Greenplate, B. J. Anderson, M. O. Harhay, T. K. Jones, J. P. Reilly, N. S. Mangalmurti, M. G. S. Shashaty, M. R. Betts, E. J. Wherry, and N. J. Meyer. Validating a proteomic signature of severe COVID-19. *Crit Care Explor*, 4(12):e0800, Dec. 2022.

R. Couillet and Z. Liao. *Random matrix methods for machine learning*. Cambridge University Press, Cambridge, England, Aug. 2022.

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

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.

E. Dobriban. Sharp detection in PCA under correlations: All eigenvalues matter. *AoS*, 45(4):1810–1833, Aug. 2017.

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.

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.

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

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

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.

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.

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.

J. H. Foley, P. Y. Kim, D. Hendriks, J. Morser, A. Gils, N. J. Mutch, and Subcommittee on Fibrinolysis. Evaluation of and recommendation for the nomenclature of the CPB2 gene product (also known as TAFI and proCPU): communication from the SSC of the ISTH. *J. Thromb. Haemost.*, 13(12):2277–2278, Dec. 2015.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1): 17–40, Jan. 1976.

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.

R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *[Revue fr. autom. inform. rech. opér., Anal. numér.*, 9(R2):41–76, 1975.

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.

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.

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.

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.

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.

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

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

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.

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.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.*, 29(2):295–327, 2001.

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.

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

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.

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.

J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, volume 5.1, pages 281–298. University of California Press, Jan. 1967.

D. A. Martino. The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659, 2014.

D. A. Martino. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data*, 4:170010, 2017.

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

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.

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

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.

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.

D. Paul. Asymptotics of sample eigenstructure for a large dimensional spiked covariance model. *Stat. Sin.*, 17(4):1617–1642, 2007.

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.

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.

P. Radchenko and G. Mukherjee. Convex clustering via l1fusion penalization. *J. R. Stat. Soc. Series B Stat. Methodol.*, 79(5):1527–1546, Nov. 2017.

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

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.

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.

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.

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.

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

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.

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.

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

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.

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

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.

M. Wang, T. Yao, and G. I. Allen. Supervised convex clustering. *Biometrics*, Mar. 2023.

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.

W. Wang, R. Arora, K. Livescu, and J. Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1083–1092. JMLR.org, July 2015.

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.

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

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.

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

J. Xie, R. Girshick, and A. Farhadi. Unsupervised deep embedding for clustering analysis. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 478–487, New York, New York, USA, 2016. PMLR.

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.

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, 2016a.

C. You, D. Robinson, and R. Vidal. Scalable sparse subspace clustering by orthogonal matching pursuit. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3918–3927, 2016b.

Y. Zhang, K. Han, C. Du, R. Li, J. Liu, H. Zeng, L. Zhu, and A. Li. Carboxypeptidase B blocks ex vivo activation of the anaphylatoxin-neutrophil extracellular trap axis in neutrophils from COVID-19 patients. *Crit. Care*, 25(1):51, Feb. 2021.

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.

Checklist

1. For all models and algorithms presented, check if you include:

(a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

(b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

(c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]

2. For any theoretical claim, check if you include:

(a) Statements of the full set of assumptions of all theoretical results. [Yes]

(b) Complete proofs of all theoretical results. [Yes]

(c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

(a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

(b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]

(c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]

(d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

(a) Citations of the creator If your work uses existing assets. [Yes]

(b) The license information of the assets, if applicable. [Not Applicable]

(c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

(d) Information about consent from data providers/curators. [Yes]

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix: Simple and Scalable Algorithms for Cluster-Aware Precision Medicine

Amanda M. Buch
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine,
Cornell University
amb2022@med.cornell.edu

Conor Liston
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine,
Cornell University
col2004@med.cornell.edu

Logan Grosenick
Dept. of Psychiatry & BMRI,
Weill Cornell Medicine
Cornell University
log4002@med.cornell.edu

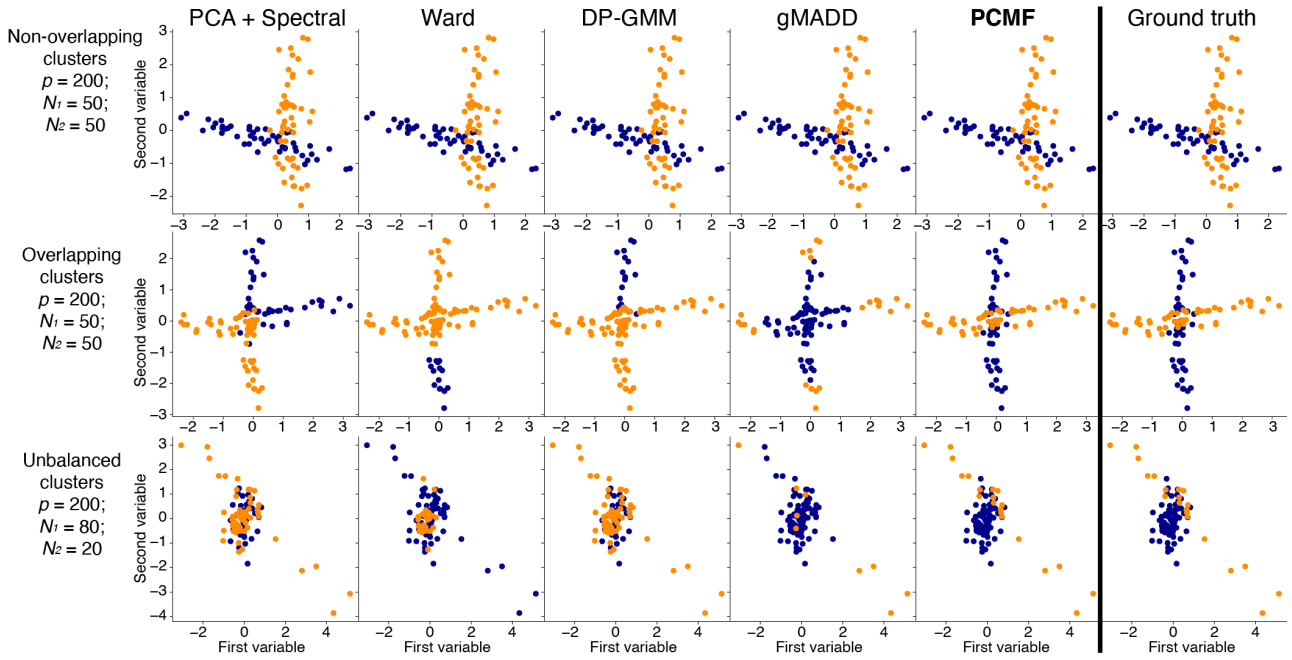

Figure 1: Examples of the three types of synthetic datasets considered in Appendix Table 1 for comparing other standard clustering methods and PCMF. Note: Sample sizes and means are slightly different than in Appendix Table 1, but the distributions are representative. Points are colored by class labels predicted by clustering method.

1 SOCIETAL IMPACT

We introduce a new interpretable method to decompose latent structure in clustered data, which has many potential positive applications for society, including as a tool to decode multiomic and neurobiological heterogeneity underlying medical disorders such as cancer and neuropsychiatric disorders. Our method is also general in that it can be applied to different types of data to learn latent covariance structure that exists within clusters of single-view or multiview data. As with any clustering algorithm, the potential for negative societal impact depends on the dataset and the intended use of the analysis. It will be critical for the use of such clustering algorithms to be on datasets that following ethical guidelines in terms of data collection and data use. The robustness and explainability provided by our methods for unbalanced, small clusters could enable positive or negative impacts on underrepresented classes depending on its application.

Contents

List of Algorithms

2 PROBLEM FORMULATION, DERIVATIONS, AND ALGORITHMS

2.1 Pathwise Clustered Matrix Factorization (PCMF)

Derivation:

Denote the PCMF optimization problem for truncated SVD of rank r:

$$\begin{aligned} \underset{\widehat{X},G,U_r,S_r,V_r}{\text{minimize}} \ & \frac{1}{2}\|X-\widehat{X}\|_F^2 + \lambda P_{\mathbf{w},q}(G) \\ \text{subject to } & \widehat{X} - U_r S_r V_r^T = 0,\ U_r^T U_r = V_r^T V_r = I_r, \\ & S_r = \text{diag}(s_1,\ldots,s_r),\ s_1 \geq \ldots \geq s_r > 0, \\ & G - D\widehat{X} = 0. \end{aligned} \tag{1}$$

For simplicity we suppress the rank subscripts letting $U = U_r,\ S = S_r,\ V = V_r$ in the following. We use the multi-convex scaled form of ADMM (Boyd et al., 2011a), yielding ADMM updates from the augmented Lagrangian with penalty parameter $\rho > 0$:

$$\begin{aligned} \widehat{X}^{k+1} \leftarrow \underset{\widehat{X}\in\mathbb{R}^{N\times p}}{\text{argmin}} \ & \frac{1}{2}\|X-\widehat{X}\|_F^2 + \frac{\rho}{2}\|U^k S^k V^k - Z_2^k\|_F^2 \\ & + \frac{\rho}{2}\|D\widehat{X} - G^k - Z_1^k\|_F^2, \end{aligned} \tag{2}$$

$$G^{k+1} \leftarrow \underset{G\in\mathbb{R}^{|\mathcal{E}|\times p}}{\text{argmin}} \ \lambda P_{\mathbf{w},q}(G) + \frac{\rho}{2}\|D\widehat{X}^{k+1} - G + Z_1^k\|_F^2, \tag{3}$$

$$(U^{k+1},S^{k+1},V^{k+1}) \leftarrow \underset{\substack{U\in\mathbb{R}^{N\times r},V\in\mathbb{R}^{p\times r}, \\ U^TU=V^TV=I_r, \\ S=\text{diag}(s_1,\ldots,s_r),\ s_1\geq\cdots\geq s_r>0}}{\text{argmin}} \|\widehat{X}^{k+1} - USV^T + Z_2^k\|_F^2, \tag{4}$$

$$Z_1^{k+1} \leftarrow Z_1^k + D\widehat{X}^{k+1} - G^{k+1}, \tag{5}$$

$$Z_2^{k+1} \leftarrow Z_2^k + \widehat{X}^{k+1} - U^{k+1}S^{k+1}V^{k+1,T}. \tag{6}$$

Note that the G^{k+1} update can be re-expressed using the proximal operator for $P_{\mathbf{w},q}(G)$, since

$$\begin{aligned} \underset{G\in\mathbb{R}^{|\mathcal{E}|\times p}}{\text{argmin}} \ \lambda P_{\mathbf{w},q}(G) + \frac{\rho}{2}\|D\widehat{X}^{k+1} - G + Z_1^k\|_F^2 &= \underset{G\in\mathbb{R}^{|\mathcal{E}|\times p}}{\text{argmin}} \ \frac{\lambda}{\rho}P_{\mathbf{w},q}(G) + \frac{\rho}{2}\left\|G - \left(D\widehat{X}^{k+1} + Z_1^k\right)\right\|_F^2 \\ &= \text{prox}_{\frac{\lambda}{\rho}P_{\mathbf{w},q}(G)}\left(D\widehat{X}^{k+1} + Z_1^k\right). \end{aligned} \tag{7}$$

Further, due to the Eckhart-Young Theorem (Eckart and Young, 1936), the solution to the $(U^{k+1},S^{k+1},V^{k+1})$ update can be expressed as simply the SVD of $\widehat{X}^{k+1} + Z_2^k$. Further, writing out the the primal update, taking the gradient with respect to $\widehat{X}$ and setting the gradient equal to zero yields updates:

$$\widehat{X}^{k+1} \ \leftarrow \ (I + \rho I + D^TD)^{-1}(X + \rho D^T(G^k - Z_1^k) + \rho(U^kS^kV^k - Z_2^k)). \tag{8}$$

Finally, letting $LL^T = I + \rho I + D^TD$, we obtain the updates in Main Text Algorithm 1 (reproduced as "Appendix Algorithm 1" here for ease of reference).

Algorithm 1 PCMF

Input: data X, path $\{\lambda\}$, weights $\mathbf{w}$, $\rho \geq 1$,
Notation: data mean $\overline{X}$, rank r, iteration k,
norm $q \in \{1, 2, \infty\}$, pairwise distance matrix D,
proximal operator of $P_{\mathbf{w},q}(\cdot)$: $\text{prox}_{\frac{\lambda}{\rho}P_{\mathbf{w},q}(\cdot)}$

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

Table 1: Clustering accuracy of PCMF on two-class data (10 runs per synthetic data type). We generate X from two distributions that differ in centroid and slope: non-overlapping clusters (centroids $\in \{-0.2, 0.2\}$), overlapping clusters (centroids $\in \{-0.05, 0.05\}$), or non-overlapping but unbalanced cluster size (centroids $\in \{-0.2, 0.2\}$). δ indicates the fraction of variables redundantly containing signal. $N.N.$ denotes number of nearest neighbors used in the convex clustering penalty.

	Non-overlapping $\delta = 0.5; N_1, N_2 = 50$		Non-overlapping $\delta = 0.2; N_1, N_2 = 50$		Overlapping $\delta = 0.5; N_1, N_2 = 50$		Overlapping $\delta = 0.2; N_1, N_2 = 50$		Unbalanced $\delta = 0.5; N_1 = 80, N_2 = 20$		Unbalanced $\delta = 0.2; N_1 = 80, N_2 = 20$	
	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI	ARI	NMI
$p = 200$												
PCA+K-means	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.06 ± 0.07	0.10 ± 0.11	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Ward	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.31 ± 0.36	0.40 ± 0.32	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Spectral	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
DP-GMM	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.25 ± 0.10	0.33 ± 0.11	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Elastic Subspace	0.16 ± 0.09	0.25 ± 0.13	0.06 ± 0.10	0.09 ± 0.14	0.80 ± 0.09	0.75 ± 0.09	0.09 ± 0.08	0.14 ± 0.11	0.05 ± 0.04	0.05 ± 0.03	0.03 ± 0.03	0.04 ± 0.03
gMADD	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.0 ± 0.01	0.02 ± 0.02	0.61 ± 0.48	0.61 ± 0.48	0.16 ± 0.01	0.27 ± 0.02	0.16 ± 0.0	0.28 ± 0.0
PCMF; No $N.N.$	0.01 ± 0.03	0.03 ± 0.04	0.0 ± 0.01	0.01 ± 0.01	0.11 ± 0.30	0.12 ± 0.30	0.0 ± 0.01	0.01 ± 0.01	0.01 ± 0.03	0.03 ± 0.04	0.0 ± 0.01	0.01 ± 0.01
LL-PCMF; No $N.N.$	0.54 ± 0.46	0.54 ± 0.45	0.60 ± 0.49	0.60 ± 0.49	0.09 ± 0.18	0.11 ± 0.18	0.14 ± 0.26	0.13 ± 0.22	0.54 ± 0.46	0.54 ± 0.45	0.60 ± 0.49	0.60 ± 0.49
PCMF; $N.N. = 25$	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**
LL-PCMF; $N.N. = 25$	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.62 ± 0.45	0.64 ± 0.42	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**
$p = 2000$												
PCA+K-means	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.05 ± 0.08	0.09 ± 0.12	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Ward	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.24 ± 0.26	0.34 ± 0.22	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Spectral	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.13 ± 0.05	0.23 ± 0.09	0.16 ± 0.0	0.28 ± 0.0
DP-GMM	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.35 ± 0.32	0.41 ± 0.30	**1.0 ± 0.0**	**1.0 ± 0.0**	0.16 ± 0.0	0.28 ± 0.0	0.16 ± 0.0	0.28 ± 0.0
Elastic Subspace	0.08 ± 0.11	0.13 ± 0.16	0.09 ± 0.12	0.14 ± 0.16	0.39 ± 0.08	0.32 ± 0.06	0.11 ± 0.06	0.20 ± 0.10	0.06 ± 0.04	0.05 ± 0.03	0.04 ± 0.04	0.04 ± 0.04
gMADD	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	0.0 ± 0.01	0.01 ± 0.02	0.80 ± 0.39	0.81 ± 0.38	0.16 ± 0.01	0.27 ± 0.02	0.16 ± 0.0	0.28 ± 0.0
PCMF; No $N.N.$	0.41 ± 0.49	0.43 ± 0.47	0.0 ± 0.0	0.0 ± 0.01	0.0 ± 0.0	0.02 ± 0.02	0.40 ± 0.49	0.40 ± 0.49	0.40 ± 0.49	0.43 ± 0.47	0.0 ± 0.01	0.01 ± 0.02
LL-PCMF; No $N.N.$	0.43 ± 0.45	0.42 ± 0.45	0.44 ± 0.46	0.44 ± 0.45	0.06 ± 0.11	0.08 ± 0.15	0.14 ± 0.28	0.14 ± 0.27	0.43 ± 0.45	0.42 ± 0.45	0.44 ± 0.46	0.44 ± 0.45
PCMF; $N.N. = 25$	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**
LL-PCMF; $N.N. = 25$	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**	**1.0 ± 0.0**

2.2 Consensus PCMF

We show the consensus formulation of PCMF in Appendix Algorithm 2 and discuss correcting for batch effects in the following subsection.

Correcting for batch effects in Consensus PCMF solution paths:
Differences in batch means result in very similar but slightly shifted paths for the different batches, an effect that results in problems for our dendrogram fitting approach. We therefore correct for the batch effects using the following procedure: (1) cluster on the last solution in the solution path (at $\lambda = 0$), (2) for each of these clusters, calculate its centroid at each time point and trace this path back to $\lambda = \infty$, (3) do this for all terminal clusters to

yield a corrected set of paths that can then be input into our dendrogram fitting procedure.

In Appendix Algorithm 2, we split X into B batches. For each batch we store the batch mean $\mu_{\mathbf{b}}$ and center X_b by subtracting $\mu_{\mathbf{b}}$. The $\widehat{X_b}^{k+1}$ and G^{k+1} updates are run on each batch X_b and then the $\widehat{X_b}^{k+1}$ across B batches are stacked prior to calculating the SVD. This yields an almost-centered $U^{k+1}, S^{k+1}, V^{k+1}$ that needs to be batch corrected by accounting for the individual batch means. We recalculate $T^{k+1} = U^{k+1}S^{k+1}$ by the dot product of the update $(\widehat{X_b}^{k+1} + Z_2{}^k)$ + the X_b mean $\mu_{\mathbf{b}}$ with the non-centered V^{k+1}. Finally, we re-calculate the U^{k+1} by dividing the centered T^{k+1} by S^{k+1} and unit normalizing the U^{k+1}. This yields a consensus U^{k+1} update that is equivalent to the centered SVD U^{k+1} update. We use this U^{k+1} for the $Z_1{}_b^{k+1}$ and $Z_1{}_b^{k+1}$ updates that we calculate in each batch. Finally, we stack the output $\widehat{X_b}^{k+1}$ across batches to yield $\widehat{X}^{k+1}$.

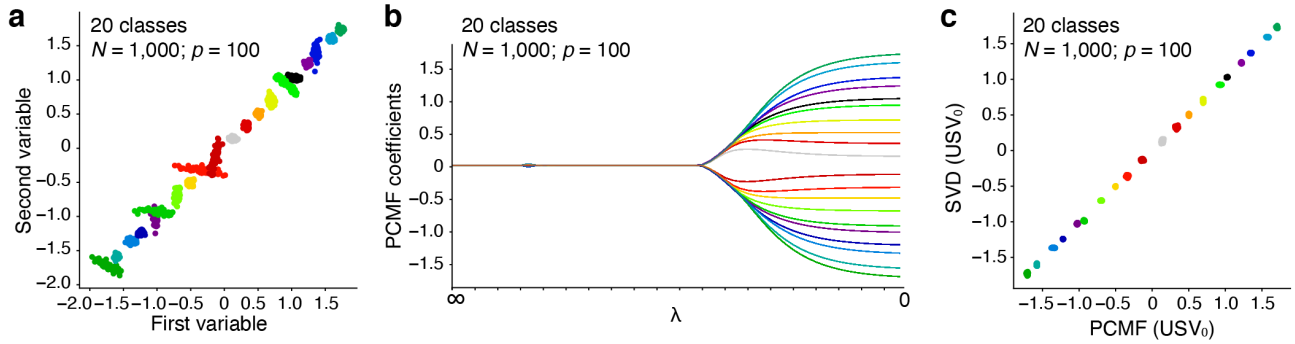

Figure 2: PCMF identifies a hierarchy of clustered matrix factorizations ($\widehat{X}$) along a convex penalty path (λ). **a.** Observations for first two variables colored by class for a synthetic dataset with 20 classes ($N = 1,000$; $p = 100$). **b.** Consensus PCMF path dendrogram of $\widehat{X}$ using 10 mini-batches of $N_{batch} = 100$. **c.** Comparison of PCMF embeddings ($r = 1$; USV_0 from PCMF) to per-cluster tSVD ($r = 1$; USV_0 from SVD). The two estimates are shown plotted against one another with colors indicating the 20 classes. We calculated the tSVD within each cluster using the ground truth clustering. The PCMF rank-one matrices per-cluster were highly similar to those found using tSVD within each ground truth cluster.

Table 2: Comparison of PCMF embeddings to per-cluster tSVD. Using the ground truth clustering, we calculated the tSVD within each cluster to find the within-cluster singular value S and singular vectors U and V. We compared these to the estimates from PCMF. S, U, and V for the 20 classes estimated by PCMF are highly similar to the S, U, and V calculated by the tSVD within ground truth clusters. This indicates that PCMF is successfully approximating the true within-cluster tSVD estimates (uncentered PCA).

Cluster	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
S (PCMF)	122.55	122.50	113.04	112.99	96.58	96.53	87.29	87.01	73.04	73.09	66.00	65.99	49.55	49.35	35.39	35.37	23.76	23.52	9.37	9.46
S (SVD)	122.55	122.51	113.05	113.00	96.59	96.54	87.23	87.03	73.04	73.09	66.01	66.00	49.59	49.38	35.41	35.38	23.87	23.56	9.42	9.54
$U_{0:20}$ (PCMF)	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14
$U_{0:20}$ (SVD)	0.14	0.18	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.14	0.13	0.14	0.14	0.14	0.14	0.14
$V_{0,0:20}$ (PCMF)	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10
$V_{0,0:20}$ (SVD)	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10	-0.10

Algorithm 2 Consensus PCMF

Input: data X, $\downarrow$ path $\{\lambda\}$, weights $\mathbf{w}$, pairwise distance matrix D
Notation: data mean $\overline{X}$, batch b, batch mean $\mu_{\mathbf{b}}$, concatenate $\|$, rank r, iteration k, norm q, $\rho > 0$
Initialize:
Split X into B batches and demean
$X = \|_{b=1}^{B} X_b \ = \overline{X_1} \,\|\, \overline{X_2} \,\|\, \cdots \,\|\, \overline{X_b}$
$\|_{b=1}^{B} \mu_b = X_b - \overline{X_b}$
$L = \text{chol}(I + \rho I + \rho D^T D)$
for $b = 1, \ldots, B$ **do**
  $G_b^0 = Z_{1b}^{\ 0} = D\overline{X_b}$
  $\widehat{X_b} = Z_{2b}^{\ 0} = \overline{X}$
end for
$({U_r}^0, {S_r}^0, {V_r}^0) \leftarrow \text{SVD}_r(\widehat{X})$
ADMM:
for $\lambda \in \{\lambda\}$ **do**
  for $k = 1, \ldots, K$ **do**
    for $b = 1, \ldots, B$ **do**
      $\widehat{X_b}^{k+1} \leftarrow L^{-T}L^{-1}\big(X_b + \rho D^T(G_b^k - Z_{1b}^{\ k}) + \rho({U_r}_b^{\ k}{S_r}^k{V_r}^{k^T} - Z_{2b}^{\ k})\big)$
      $G_b^{k+1} \leftarrow \text{prox}_{\lambda/\rho, q(\cdot,\cdot)}(D\widehat{X_b}^{k+1} + Z_{1b}^{\ k}; \mathbf{w}_b)$
    end for
    $\widehat{X}^{k+1} = \|_{b=1}^{B} \widehat{X_b}^{k+1}, \quad {Z_2}^k = \|_{b=1}^{B} Z_{2b}^{\ k}$
    $({U_b}_r^{k+1}, S_r^{k+1}, {V_b}_r^{k+1}) \leftarrow \text{SVD}_r(\widehat{X}^{k+1} + Z_2^k)$
    $T_r^{k+1} \leftarrow (\widehat{X}^{k+1} + Z_2^k + \mu_b){V_b}_r^{k+1}$
    $U_r^{k+1} \leftarrow T_r^{k+1}/S_r^{k+1}$
    $(U_r^{k+1})_{\cdot j} \ = \frac{(U_r^{k+1})_{\cdot j}}{\|(U_r^{k+1})_{\cdot j}\|_2^2}$
    for $b = 1, \ldots, B$ **do**
      $Z_{1b}^{\ k+1} \leftarrow Z_{1b}^{\ k} + D^T\widehat{X_b}^{k+1} - G_b^{k+1}$
      $Z_{2b}^{\ k+1} \leftarrow Z_{2b}^{\ k} + \widehat{X_b}^{k+1} - {U_r}_b^{k+1}, {S_r}^{k+1}, {V_r}^{k+1}$
    end for
  end for
  $\widehat{X}_\lambda \leftarrow \widehat{X}^K = \|_{b=1}^{B} \widehat{X_b}^K, \quad G_\lambda \leftarrow G^K = \|_{b=1}^{B} G_b^K$
  $(U_{r,\lambda}, S_{r,\lambda}, V_{r,\lambda}) \leftarrow (U_r^K, S_r^K, V_r^K)$
  for $b = 1, \ldots, B$ **do**
    $\widehat{X_b}^0 \leftarrow \widehat{X_b}^K$, $G_b^0 \leftarrow G_b^K$
  end for
  $({U_r}^0, {S_r}^0, {V_r}^0) \leftarrow ({U_r}^K, {S_r}^K, {V_r}^K)$
end for

return $\{\widehat{X}_\lambda\}, \{G_\lambda\}, \{U_{r,\lambda}\}, \{S_{r,\lambda}\}, \{V_{r,\lambda}\}$

2.3 Locally Linear (LL)-PCMF

Algorithm 3 LL-PCMF

Input: data X, decreasing path $\{\lambda\}$, weights $\mathbf{w}$
Notation: data mean $\overline{X}$, samples N, iteration k, norm $q \in \{1, 2, \infty\}$, $\rho \geq 1$

$V^0 \leftarrow \overline{X}, \mathbf{s}^0 = \mathbf{1}$
for $\lambda \in \{\lambda\}$ **do**
  for $k = 1, \ldots, K$ **do**
    $y^k_{\mathbf{u},i} = \mathbf{x}_i^T \mathbf{v}_i^k, i = 1, \ldots, N$
    $\mathbf{u}^{k+\frac{1}{2}} \leftarrow \textsc{ConvexCluster}(\mathbf{y}^k_{\mathbf{u}}, \mathbf{u}^k, q)$
    $\mathbf{u}^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}^{k+\frac{1}{2}})$
    $\mathbf{y}^{k+1}_{\mathbf{v},i} = u_i^{k+1} \mathbf{x}_i, \ i = 1, \ldots, N$
    $\{\mathbf{v}_i\}^{k+\frac{1}{2}} \leftarrow \textsc{ConvexCluster}(Y^{k+1}_{\mathbf{v}}, V^k, \lambda, \mathbf{w}, q)$
    $\{\mathbf{v}_i\}^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}}), \ i = 1, \ldots, N$
  end for
  $s_i^K \leftarrow u_i^K \mathbf{x}_i^T \mathbf{v}_i^K$
  Save current path solutions: $(\mathbf{u}_\lambda, s_\lambda, V_\lambda) \leftarrow (\mathbf{u}^K, s^K, V^K)$
  Initialize for next path solution: $(\mathbf{u}^0, s^0, V^0) \leftarrow (\mathbf{u}^K, s^K, V^K)$
end for
return pathwise solutions $\{\mathbf{u}_\lambda\}, \{s_\lambda\}, \{V_\lambda\}$

Derivation and the rank-1 problem:

We present a Penalized Alternating Least Squares (PALS) approach that provides a local linear fitting of per-cluster factors. Without loss of generality, we consider the rank-1 version of this problem. Note this can be generalized to rank-r using our orthogonalized deflation procedure described in Appendix §2.5, or another an appropriate deflation approach (Mackey, 2008, Witten et al., 2009).

Consider the PCMF optimization problem:

$$\begin{aligned} \underset{\widehat{X}, U_r, S_r, V_r}{\text{minimize}} \ & \frac{1}{2}\|X - \widehat{X}\|_F^2 + \lambda \sum_{i<j} w_{ij} \|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q \\ \text{subject to } & \widehat{X} - U_r S_r V_r^T = 0, \\ & U_r^T U_r = V_r^T V_r = I_r, \ S_r = \text{diag}(s_1, \ldots, s_r), \\ & s_1 \geq s_2 \geq \cdots \geq s_r > 0, \end{aligned} \tag{9}$$

we note that the first part of the objective can be expanded as:

$$\begin{aligned} \frac{1}{2}\|X - \widehat{X}\|_F^2 &= \frac{1}{2}\text{tr}\left((X - \widehat{X})^T (X - \widehat{X})\right) \\ &= \frac{1}{2}\|X\|_F^2 - \text{tr}\left(\widehat{X}^T X\right) + \frac{1}{2}\|\widehat{X}\|_F^2 \\ &= \frac{1}{2}\|X\|_F^2 - \text{tr}\left(V_r S_r U_r^T X\right) + \frac{1}{2}\text{tr}\left(V_r S_r U_r^T U_r S_r V_r^T\right) \\ &= \frac{1}{2}\|X\|_F^2 - \text{tr}\left(S_r U_r^T X V_r\right) + \frac{1}{2}\sum_{i=1}^r s_i^2 \\ &= \frac{1}{2}\|X\|_F^2 - \sum_{i=1}^r s_i \mathbf{u}_i^T X \mathbf{v}_i + \frac{1}{2}\sum_{i=1}^r s_i^2, \end{aligned} \tag{10}$$

where $\mathbf{v}_r$ and $\mathbf{u}_r$ are column vectors, and that the final two equivalences use the fact that $U^T U = I_r, \ V^T V = I_r$. In the rank-1 case ($r = 1$), and letting $\mathbf{u} = \mathbf{u}_1, \ \mathbf{v} = \mathbf{v}_1$, and denoting the first singular value s, this can be

rewritten:

$$\begin{aligned} &\underset{\widehat{X},s,\mathbf{u},\mathbf{v}}{\text{minimize}} \ \frac{1}{2}\|X\|_F^2 - s\mathbf{u}^T X\mathbf{v} + \frac{s^2}{2} + \lambda \sum_{i<j} w_{ij}\|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q, \\ &\text{subject to } \widehat{X}_{i\cdot} = su_i\mathbf{v}^T, \ \|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}\|_2^2 = 1, \ s > 0. \end{aligned} \tag{11}$$

By simplifying by setting $s = 1$ (see the subsection 4.3 for an approach with more general s) and considering at the gradient of the objective, we can see this problem has the same solution as the following problem (Witten et al., 2009):

$$\begin{aligned} &\underset{\widehat{X},\mathbf{u},V}{\text{minimize}} \ -\mathbf{u}^T X\mathbf{v} + \lambda \sum_{i<j} w_{ij}\|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q, \\ &\text{subject to } \widehat{X}_{i\cdot} = u_i\mathbf{v}^T, \ \|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}\|_2^2 = 1. \end{aligned} \tag{12}$$

Augmenting the rank-1 problem for PCMF:

However, this formulation does not allow our rank-1 problem to approximate every element of X with a corresponding element of $\widehat{X}$ as required by the convex clustering formulation. To remedy this, we introduce the overparameterization, letting $\mathbf{v}_i = V_{\cdot i}$ (the ith column of matrix $V \in \mathbb{R}^{p\times N}$):

$$\begin{aligned} &\underset{\widehat{X},\mathbf{u},V}{\text{minimize}} \ \sum_{i=1}^{N} -u_i X_{i\cdot}\mathbf{v}_i + \lambda \sum_{i<j} w_{ij}\|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q, \\ &\text{subject to } \widehat{X}_{i\cdot} = u_i\mathbf{v}_i, \ \|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}_i\|_2^2 = 1, \ i = 1,\dots,N, \end{aligned} \tag{13}$$

where each row of X is now approximated by a potentially unique value $\widehat{X}_{i\cdot} = u_i\mathbf{v}_i$, just as in the convex clustering problem. This overparameterization means that the solution to the first term in the optimization problem solved for PCMF is able to fit each data point exactly, except to the extent that it is limited by the penalty term. This leads to the nearest-neighbor or kernel-based weights in the penalty term determining the locally linear approximation to the data.

Next, letting $\mathbf{x}_i = (X_{i\cdot})^T$ (a column of X^T), we note that due to the quadratic equality constraints $\|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}_i\|_2^2 = 1, \ i = 1,\dots,N$, the following relationships hold:

$$\begin{aligned} \frac{1}{2}\sum_{i=1}^{N}\|\mathbf{x}_i^T\mathbf{v}_i - u_i\| &= \frac{1}{2}\sum_{i=1}^{N}\mathbf{v}_i^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_i - \sum_{i=1}^{N}u_i\mathbf{x}_i^T\mathbf{v}_i + \sum_{i=1}^{N}u_i^2 \\ &= \frac{1}{2}\sum_{i=1}^{N}\mathbf{v}_i^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}_i - \sum_{i=1}^{N}u_i\mathbf{x}_i^T\mathbf{v}_i + 1, \end{aligned} \tag{14}$$

and

$$\frac{1}{2}\|u_i\mathbf{x}_i - \mathbf{v}_i\| = \frac{1}{2}u_i^2\mathbf{x}_i^T\mathbf{x}_i - u_i\mathbf{x}_i^T\mathbf{v}_i + \mathbf{v}_i^T\mathbf{v}_i = -u_i\mathbf{x}_i^T\mathbf{v}_i + \frac{1}{2}\mathbf{x}_i^T\mathbf{x}_i + 1, \tag{15}$$

and therefore it follows that (letting $y_{\mathbf{u},i}^k = \mathbf{x}_i^T\mathbf{v}_i^k$ and $\mathbf{y}_{\mathbf{v},i}^k = u_i^k\mathbf{x}_i$) the updates:

$$\mathbf{u}^{k+1} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \ \sum_{i=1}^{N} -u_i\mathbf{x}_i^T\mathbf{v}_i^k + \lambda P_{\mathbf{w},q}(\mathbf{u}, V^k), \text{ subject to } \|\mathbf{u}\|_2^2 = 1, \tag{16}$$

$$\begin{aligned} \{\mathbf{v}_i\}^{k+1} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \ &\sum_{i=1}^{N} -u_i^{k+1}\mathbf{x}_i^T\mathbf{v}_i + \lambda P_{\mathbf{w},q}(\mathbf{u}^{k+1}, V), \\ &\text{subject to } \|\mathbf{v}_i\|_2^2 = 1, i = 1,\dots,N, \end{aligned} \tag{17}$$

have the same solutions as the updates:

$$\mathbf{u}^{k+1} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \sum_{i=1}^{N} \|y^k_{\mathbf{u},i} - u_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}, V^k), \text{ subject to } \|\mathbf{u}\|_2^2 = 1, \tag{18}$$

$$\{\mathbf{v}_i\}^{k+1} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^{N} \|\mathbf{y}^{k+1}_{\mathbf{v},i} - \mathbf{v}_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}^{k+1}, V), \tag{19}$$
$$\text{subject to } \|\mathbf{v}_i\|_2^2 = 1, i = 1, \ldots, N.$$

To enforce the quadratic equality constraints, we use proximal projection (Parikh and Boyd, 2014) updates following K iterations of the $\mathbf{u}$ and $\{\mathbf{v}_i\}$ updates, projecting onto the squared L2 unit ball associated with the equality constraints as intermediate steps in the algorithm. Thus if $k = 1$ we would have:

$$\mathbf{u}^{k+\frac{1}{2}} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \sum_{i=1}^{N} \|y^k_{\mathbf{u},i} - u_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}, V^k), \tag{20}$$

$$\mathbf{u}^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}^{k+\frac{1}{2}}), \tag{21}$$

$$\{\mathbf{v}_i\}^{k+\frac{1}{2}} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^{N} \|\mathbf{y}^{k+1}_{\mathbf{v},i} - \mathbf{v}_i\|_2^2 + \lambda P_{\mathbf{w},q}(\mathbf{u}^{k+1}, V), \tag{22}$$

$$\{\mathbf{v}_i\}^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}}), \; i = 1, \ldots, N, \tag{23}$$

where

$$\text{prox}_{\|\cdot\|_2^2}(\mathbf{a}) = \begin{cases} \frac{\mathbf{a}}{\|\mathbf{a}\|_2^2} & \text{if } \|\mathbf{a}\|_2^2 > 1 \\ \mathbf{a} & \text{if } \|\mathbf{a}\|_2^2 \leq 1. \end{cases} \tag{24}$$

To decrease computation time and significantly increase the flexibility of fitting per-cluster factors with $r = 1$, we remove the cross terms in the penalty; we relax the problem formulation by replacing $P_{\mathbf{w},q}(\mathbf{u}, \mathbf{v})$ with $Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u}) = \sum_{(i,j)\in\mathcal{E}} w_{ij}|u_i - u_j|$ and $Q^{V}_{\mathbf{w},q}(V) = \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_q$.

This yields updates:

$$\mathbf{u}^{k+1} \leftarrow \underset{\mathbf{u}}{\text{argmin}} \sum_{i=1}^{N} \|y^k_{\mathbf{u},i} - u_i\|_2^2 + \lambda Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u}), \text{ subject to } \|\mathbf{u}\|_2^2 = 1, \tag{25}$$

$$\{\mathbf{v}_i\}^{k+1} \leftarrow \underset{\{\mathbf{v}_i\}}{\text{argmin}} \sum_{i=1}^{N} \|\mathbf{y}^{k+1}_{\mathbf{v},i} - \mathbf{v}_i\|_2^2 + \lambda Q^{V}_{\mathbf{w},q}(V), \text{ subject to } \|\mathbf{v}_i\|_2^2 = 1, \; i = 1, \ldots, N. \tag{26}$$

We find that by relaxing $P_{\mathbf{w},q}(\mathbf{u}, V)$ to $Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u})$ and $Q^{V}_{\mathbf{w},q}(V)$ in our iterations, there is no need to recompute the difference matrix for each penalty, allowing the Cholesky factorization associated with that matrix to be cached to speed up computations (often significantly) and still yielding good clustering performance (see experiments in Appendix Table 1).

Both updates are standard convex clustering problems that are solvable in a number of ways. Here we use ADMM updates that can be easily incorporated into an Algorithmic Regularization scheme to speed up the rate of convergence in the optimization problem (described in Appendix §2.7 below). We use the updates from (Weylandt et al., 2020) (see their Appendix Algorithms 1 and 2), and we will subsequently refer to this convex clustering solver algorithm as ConvexCluster (rewritten in our notation in Appendix Algorithms 4 and 5). Here we implement ConvexCluster as an ADMM solver that can be scaled using consensus, but other convex clustering solvers could be used instead. Putting these together, we obtain Appendix Algorithm 3. Finally, for additional intuition as to why such multi-block algorithms may be advantageous, see (Goncalves et al., 2019), although note that when fit along the regularization path the PCMF objective is not guaranteed to monotonically decrease (see Appendix Fig. 4a).

Allowing different singular values for different clusters in rank-1 LL-PCMF:

Above we have glossed over an important detail regarding the fitting of LL-PCMF singular values.

Somewhat more flexibly than the linear PCMF algorithm, the LL-PCMF relaxation allows both $\mathbf{u}$ and $\{\mathbf{v}_i\}$ to vary quite flexibly such that distinct clusters may be fit easily when $r = 1$. This highlights the question (also of relevance in the linear PCMF case, but perhaps less obviously): what if distinct clusters have different singular values in this rank-1 LL-PCMF case?

One possible solution to this problem is the following. Returning to our overparameterization where we set $\mathbf{v}_i = V_{\cdot i}$ (for $V \in \mathbb{R}^{p \times N}$), we now also overparameterize the singular value, letting each observation have its own s_i for $i = 1, \ldots, N$. From above (noting that now $\widehat{X}_{i\cdot} = s_i u_i \mathbf{v}_{i\cdot}^T$), we would like the following minimization problem to be solved:

$$\begin{aligned} &\underset{\mathbf{s},\mathbf{u},V}{\text{minimize}} \ \sum_i -s_i u_i X_{i\cdot}\mathbf{v}_i + \frac{1}{2}\sum_i s_i^2 + \lambda \sum_{i<j} w_{ij} \|s_i u_i \mathbf{v}_{i\cdot} - s_j u_j \mathbf{v}_{j\cdot}\|_q, \\ &\text{subject to} \ \ \|\mathbf{u}\|_2^2 = 1, \ \|\mathbf{v}_i\|_2^2 = 1, \ i = 1, \ldots, N. \end{aligned} \tag{27}$$

If we continue to use the relaxed penalties $Q^{\mathbf{u}}_{\mathbf{w},q}(\mathbf{u}) = \sum_{(i,j)\in\mathcal{E}} w_{ij}|u_i - u_j|$ and $Q^{V}_{\mathbf{w},q}(V) = \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\mathbf{v}_i - \mathbf{v}_j\|_q$ as above (allowing them to not depend on the s_i), then we can ignore the penalty term in the above problem and, given fixed estimates $\mathbf{u}^K$ and $\{\mathbf{v}_i\}^K$, instead solve:

$$\underset{\mathbf{s}}{\text{minimize}} \ \sum_i -s_i u_i^K X_{i\cdot}\mathbf{v}_i^K + \frac{1}{2}\sum_i s_i^2 \ . \tag{28}$$

Taking the gradient of the objective with respect to $\mathbf{s}$, setting it equal to zero, and solving we then obtain the solution (leading to line 11 of Appendix Algorithm 3):

$$s_i^K \leftarrow u_i^K \mathbf{x}_i^T \mathbf{v}_i^K \tag{29}$$

Algorithm 4 ConvexCluster($\mathbf{y}^k_{\mathbf{u},\lambda}, \mathbf{u}^k_\lambda, q$): $\mathbf{u}$ Update

Input: Auxiliary variable $\mathbf{y}^k_{\mathbf{u}}$, previous iterate $\mathbf{u}^k$, norm q
Notation: pairwise distance matrix D, iteration k, $\rho \geq 1$
Initialize (if k = 0):
$L = \text{chol}(I + \rho I + \rho D^T D)$
$W_2 = Z_2 = D{\mathbf{y}_{\mathbf{u}}}^0$
ADMM:
$\mathbf{u}^{k+\frac{1}{2}} = L^{-T}L^{-1}\big({\mathbf{y}_{\mathbf{u}}}^k + \rho D^T({W_2}^k - {Z_2}^k)\big)$
${W_2}^{k+1} = \text{prox}_{\lambda/\rho,\|\cdot\|_2^2}(D\mathbf{u}^{k+\frac{1}{2}} + {Z_2}^k)$
${Z_2}^{k+1} = {Z_2}^k + D\mathbf{u}^{k+\frac{1}{2}} - {W_2}^{k+1}$
return $\mathbf{u}^{k+\frac{1}{2}}_\lambda$

Algorithm 5 ConvexCluster($Y^{k+1}_{\mathbf{v},\lambda}, V^k_\lambda, \lambda, \mathbf{w}, q$): V Update

Input: Auxiliary variable $Y^{k+1}_{\mathbf{v}}$, previous iterate V^k, penalty λ, weights $\mathbf{w}$, norm q
Notation: pairwise distance matrix D, iteration k, $\rho \geq 1$
Initialize (if k = 0):
$L = \text{chol}(I + \rho I + \rho D^T D)$
${W_1} = Z_1 = DX$
ADMM:
$V^{k+\frac{1}{2}} = L^{-T}L^{-1}\big(Y^{k+1}_{\mathbf{v}} + \rho D^T({W_1}^k - {Z_1}^k)\big)$
${W_1}^{k+1} = \text{prox}_{\lambda/\rho,\|\cdot\|_2^2}(DV^{k+\frac{1}{2}} + {Z_1}^k)$
${Z_1}^{k+1} = {Z_1}^k + DV^{k+\frac{1}{2}} - {W_1}^{k+1}$
return $V^{k+\frac{1}{2}}_\lambda$

ConvexCluster algorithm:

We remark in the manuscript that to solve the convex clustering problem we use the ADMM approach of Weylandt et al. (2020) — see their Appendix Algorithms A1 and A2 for details and our Appendix Algorithms 4 and 5 below for use. We choose this ADMM approach in particular as it is efficient and amenable to running for just a few ADMM iterations at each step inside our own iterative algorithms, allowing us to apply Algorithmic Regularization along the path of solutions.

2.4 Pathwise Clustered Canonical Correlation Analysis (P3CA)

Algorithm 6 Pathwise Clustered Canonical Correlation Analysis (P3CA)

Input: data (X, Y), path $\{\lambda\}$, weights $\mathbf{w}$, $\rho \geq 1$,
Notation: iter. k, data means $(\bar{X}, \bar{Y})$, $\mathbf{v}_i = V_{\cdot i}$,
$\tilde{\mathbf{x}}_i = (\tilde{X}_{i\cdot})^T$, $\tilde{\mathbf{y}}_i = (\tilde{Y}_{i\cdot})^T$, norm $q \in \{1, 2, \infty\}$
$U \leftarrow \bar{X}, V \leftarrow \bar{Y}$
for $\lambda \in \{\lambda\}$ **do**
  for $k = 1, \ldots, K$ **do**
    $\tilde{\mathbf{x}}_i^{k+1} \leftarrow \Sigma_i \mathbf{v}_i^k$ $(\Sigma_i = X_{i\cdot} Y_{i\cdot}^T \in \mathbb{R}^{p_X \times p_Y})$ for $i = 1, \ldots, N$
    $\mathbf{u}_i^{k+\frac{1}{2}} \leftarrow$ ConvexCluster$(\tilde{X}^{k+1}, U^k, \lambda, \mathbf{w}, \text{q})$
    $\mathbf{u}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}_i^{k+\frac{1}{2}})$ for $i = 1, \ldots, N$
    $\tilde{\mathbf{y}}_i^{k+1} \leftarrow \Sigma_i^T \mathbf{u}_i^{k+1}$ $(\Sigma_i^T = Y_{i\cdot} X_{i\cdot}^T \in \mathbb{R}^{p_Y \times p_X})$ for $i = 1, \ldots, N$
    $\mathbf{v}_i^{k+\frac{1}{2}} \leftarrow$ ConvexCluster$(\tilde{Y}^{k+1}, V^k, \lambda, \mathbf{w}, \text{q})$
    $\mathbf{v}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}})$ for $i = 1, \ldots, N$
  end for
  Save path solutions: $U_{i\cdot}^K \leftarrow \mathbf{u}_i^{KT}$; $V_{i\cdot}^K \leftarrow \mathbf{v}_i^{KT}$ for $i = 1, \ldots, N$; $(U_\lambda, V_\lambda) \leftarrow (U^K, V^K)$
  Initialize: $(U^0, V^0) \leftarrow (U^K, V^K)$
end for
return pathwise solutions $\{U_\lambda\}, \{V_\lambda\}$

Following previous work in $p > N$ problems, we treat the covariance matrices in this problem as diagonal (Dudoit et al., 2002, Tibshirani et al., 2003, Witten et al., 2009) , and let $\Sigma = X^T Y$ resulting in the simplified problem:

$$\underset{\mathbf{u},\,\mathbf{v}}{\text{maximize}}\ \mathbf{u}^T \Sigma \mathbf{v} \text{ subject to } \|\mathbf{u}\|_2^2 = 1,\ \|\mathbf{v}\|_2^2 = 1. \tag{30}$$

In order to generalize this to convex clustering, we once again must introduce an overparameterization to allow (when $\lambda \to 0$) one unique parameter for each element of the matrix we are trying to approximate (in this case Σ). We do this by constructing the outer product matrices of the rows of X and Y as $\Sigma_i = X_{i\cdot}^T Y_{i\cdot} \in \mathbb{R}^{p_X \times p_Y}$, and then denoting the vectors $\mathbf{u}_i = (U_{i\cdot})^T$ and $\mathbf{v}_i = (V_{i\cdot})^T$ and penalty function $Q_{\mathbf{w},q}(A) = \sum_{(i,j)\in\mathcal{E}_A} w_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|_q$, we define the P3CA problem as:

$$\begin{aligned} &\underset{U,V}{\text{maximize}}\ \sum_{i=1}^N \mathbf{u}_i^T \Sigma_i \mathbf{v}_i - \lambda Q_{\mathbf{w},q}(U) - \lambda Q_{\mathbf{w},q}(V), \\ &\text{subject to } \|\mathbf{u}_i\|_2^2 = 1,\ \|\mathbf{v}_i\|_2^2 = 1,\ i = 1, \ldots, N. \end{aligned} \tag{31}$$

Following similar logic as we did for Appendix Algorithm 3 above yields updates:

$$\tilde{\mathbf{x}}_i^{k+1} \leftarrow \Sigma_i \mathbf{v}_i^k \ (\Sigma_i = X_{i\cdot}^T Y_{i\cdot} \in \mathbb{R}^{p_X \times p_Y}) \text{ for } i = 1, \dots, N, \tag{32}$$

$$\mathbf{u}_i^{k+\frac{1}{2}} \leftarrow \underset{\mathbf{u}_i}{\text{argmin}} \sum_{i=1}^{N} \|\tilde{\mathbf{x}}_i^{k+1} - \mathbf{u}_i\|_2^2 + \lambda Q_{\mathbf{w},q}(U) \text{ for } i = 1, \dots, N, \tag{33}$$

$$\mathbf{u}_i^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{u}_i^{k+\frac{1}{2}}) \text{ for } i = 1, \dots, N, \tag{34}$$

$$\tilde{\mathbf{y}}_i^{k+1} \leftarrow \Sigma_i^T \mathbf{u}_i^{k+1} \ (\Sigma_i^T = Y_{i\cdot} X_{i\cdot}^T \in \mathbb{R}^{p_Y \times p_X}) \text{ for } i = 1, \dots, N, \tag{35}$$

$$\mathbf{v}_i{}^{k+\frac{1}{2}} \leftarrow \underset{\mathbf{v}_i}{\text{argmin}} \sum_{i=1}^{N} \|\tilde{\mathbf{y}}_i^{k+1} - \mathbf{v}_i\|_2^2 + \lambda Q_{\mathbf{w},q}(V) \text{ for } i = 1, \dots, N, \tag{36}$$

$$\mathbf{v}_i{}^{k+1} \leftarrow \text{prox}_{\|\cdot\|_2^2}(\mathbf{v}_i^{k+\frac{1}{2}}) \text{ for } i = 1, \dots, N, \tag{37}$$

and once again noting that the $\mathbf{u}_i^{k+\frac{1}{2}}$ and $\mathbf{v}_i^{k+\frac{1}{2}}$ updates are convex clustering problems solvable using plugin algorithm ConvexCluster (see Appendix §2.3 and Appendix Algorithm 5), we arrive at the updates in Main Text Algorithm 2 (reproduced here as Appendix Algorithm 6 for reference).

2.5 Orthogonalized Deflation for LL-PCMF and P3CA

LL-PCMF and P3CA: Solving for rank $r > 1$ via orthogonalized deflation. To solve rank $r > 1$ LL-PCMF and P3CA problems, one can use a sequential rank-1 embedding and orthogonalized deflation with renormalization procedure (see Appendix Algorithms 7 and 8 and the LL-PCMF example in Appendix Fig. 3). We build on work from Kruger and Joe Qin (2003) (see §2.2, Eq. 6-9 for the orthogonalized deflation procedure they use to solve partial least squares) and Grosenick et al. (2013) for renormalization (see page 311 in § "Rescaling coefficients to account for 'double shrinking'"). Alternatively, other appropriate deflation procedures could be implemented, such as those outlined in Mackey (2008) or Witten et al. (2009). Note this procedure requires specifying which λ to use for specifying the PCMF/P3CA coefficients used to project the data for the PCMF/P3CA scores, which we choose here as the last (smallest) λ along the path.

Defining the rank-r PCMF score as (note here we use r to indicate the rth rank):

$$X_{PC,r} = \mathbf{u}_{\lambda,r} \mathbf{s}_{\lambda,r} = X V_{\lambda,r}, \tag{38}$$

where $\mathbf{u}_{\lambda,r}$ and $V_{\lambda,r}$, are the first r PCMF coefficients in their columns at path penalty, λ, and $X_{PC,r}$ is the rank-r PCMF score of the data matrix X, the $r+1$st deflation is given by:

$$X_{r+1} = (I - P_{\mathbf{u}_{\lambda,r}})\ X = X - X_{PC,r}\ (X_{PC,r}{}^T X_{PC,r})^{-1}\ X_{PC,r}{}^T\ X, \tag{39}$$

where $P_{\mathbf{u}_{\lambda,r}} = X_{PC,r}\ (X_{PC,r}^T X_{PC,r})^{-1}\ X_{PC,r}{}^T$ is the usual orthogonal projection matrix on the rank-r subspace of the column space of $X_{PC,r}$.

As the convex clustering penalty may cause coefficient shrinkage, we also modify these to "un-shrink" each deflation estimate by estimate scalar coefficients $\beta_{\mathbf{u}_{\lambda,r}}$ that returns the shrunk rank-r projection of the data to the same scale as the data, such that:

$$X_{r+1} = (I - P_{\mathbf{u}_{\lambda,r}})\ X = X - \beta_{\mathbf{u}(\lambda,r)}\ X_{PC,r}\ (X_{PC,r}{}^T X_{PC,r})^{-1}\ X_{PC,r}{}^T\ X, \tag{40}$$

where the coefficient $\beta_{\mathbf{u}_{\lambda,r}}$ is obtained using univariate regression model:

$$X = \beta_{\mathbf{u}_{\lambda,r}}\ P_{\mathbf{u}_{\lambda,r}}\ X + \epsilon_{\mathbf{u}_{\lambda,r}}. \tag{41}$$

For P3CA, we perform this procedure analogously on the data matrices, X and Y, with respect to the P3CA scores for X and Y (X and Y projected into the P3CA embedding subspace using P3CA coefficients U and V

accordingly). We then follow a similar procedure for P3CA (see Appendix Algorithm 8), except define the rank-r P3CA scores pair (sometimes called "canonical variate" pair) as:

$$\begin{aligned} X_{U,r} &= X_r U_{\lambda,r} \\ Y_{V,r} &= Y_r V_{\lambda,r}, \end{aligned} \tag{42}$$

where U and V are the P3CA coefficients.

Algorithm 7 LL-PCMF rank $r > 1$ via orthogonalized deflation

Input: data X
Notation: set of path solutions along decreasing path $\{\lambda\}$, singular vectors $\mathbf{u}$ and V, singular value s, principal component for rth rank $X_{PC,r}$, vectorize [:]

$\{\mathbf{u}_{\lambda,r=1}\}, \{s_{\lambda,r=1}\}, \{V_{\lambda,r=1}\} \leftarrow$ Algorithm: LL-PCMF(X)
$X_{r=1} = X$
for $r = 1, \ldots, r_{max}$ **do**
  Set $\lambda \leftarrow \lambda_{min}$
  $X_{PC,r} = X_r V_{\lambda,r}$
  $P_{\mathbf{u}_{\lambda,r}} = X_{PC,r} \ (X_{PC,r}^T X_{PC,r})^{-1} \ {X_{PC,r}}^T$
  $X_{r+1} = X_r - P_{\mathbf{u}_{\lambda,r}} \ X_r$
  Solve $X_r[:] = \beta_{\mathbf{u}_{\lambda,r}} \ X_{r+1}[:] + \epsilon_{\mathbf{u}_{\lambda,r}}$
  $X_{r+1} = \beta_{\mathbf{u}_{\lambda,r}} \ \ X_{r+1}$
  $\{\mathbf{u}_{\lambda,r+1}\}, \{s_{\lambda,r+1}\}, \{V_{\lambda,r+1}\} \leftarrow$ Algorithm: LL-PCMF(X_{r+1})
end for
return $\{\mathbf{u}_{\lambda,r+1}\}, \{s_{\lambda,r+1}\}, \{V_{\lambda,r+1}\}$

Algorithm 8 P3CA rank $r > 1$ via orthogonalized deflation

Input:data (X, Y)
Notation: set of path solutions along decreasing path $\{\lambda\}$, coefficients U and V, P3CA scores for rth rank $(X_{U,r}, Y_{V,r})$, vectorize [:]

$\{U_{\lambda,r=1}\}, \{V_{\lambda,r=1}\} \leftarrow$ Algorithm: P3CA(X, Y)
$X_{r=1} = X$ and $Y_{r=1} = Y$
for $r = 1, \ldots, r_{max}$ **do**
  Set $\lambda \leftarrow \lambda_{min}$
  $X_{U,r} = X_r U_{\lambda,r}$
  $P_{X_{\lambda,r}} = X_{U,r} \ (X_{U,r}^T X_{U,r})^{-1} \ {X_{U,r}}^T$
  $X_{r+1} = X_r - P_{X_{\lambda,r}} X_{U,r}$
  Solve $X_r[:] = \beta_{X_{\lambda,r}} \ X_{r+1}[:] + \epsilon_{U_{\lambda,r}}$
  $X_{r+1} = \beta_{X_{\lambda,r}} \ \ X_{r+1}$
  $Y_{V,r} = Y_r V_{\lambda,r}$
  $P_{Y_{\lambda,r}} = Y_{V,r} \ (Y_{V,r}^T Y_{V,r})^{-1} \ {Y_{V,r}}^T$
  $Y_{r+1} = Y_r - P_{Y_{\lambda,r}} Y_{V,r}$
  Solve $Y_r[:] = \beta_{Y_{\lambda,r}} \ Y_{r+1}[:] + \epsilon_{V_{\lambda,r}}$
  $Y_{r+1} = \beta_{Y_{\lambda,r}} \ \ Y_{r+1}$
  $\{U_{\lambda,r+1}\}, \{V_{\lambda,r+1}\} \leftarrow$ Algorithm: P3CA(X_{r+1}, Y_{r+1})
end for
return $\{U_{\lambda,r+1}\}, \{V_{\lambda,r+1}\}$

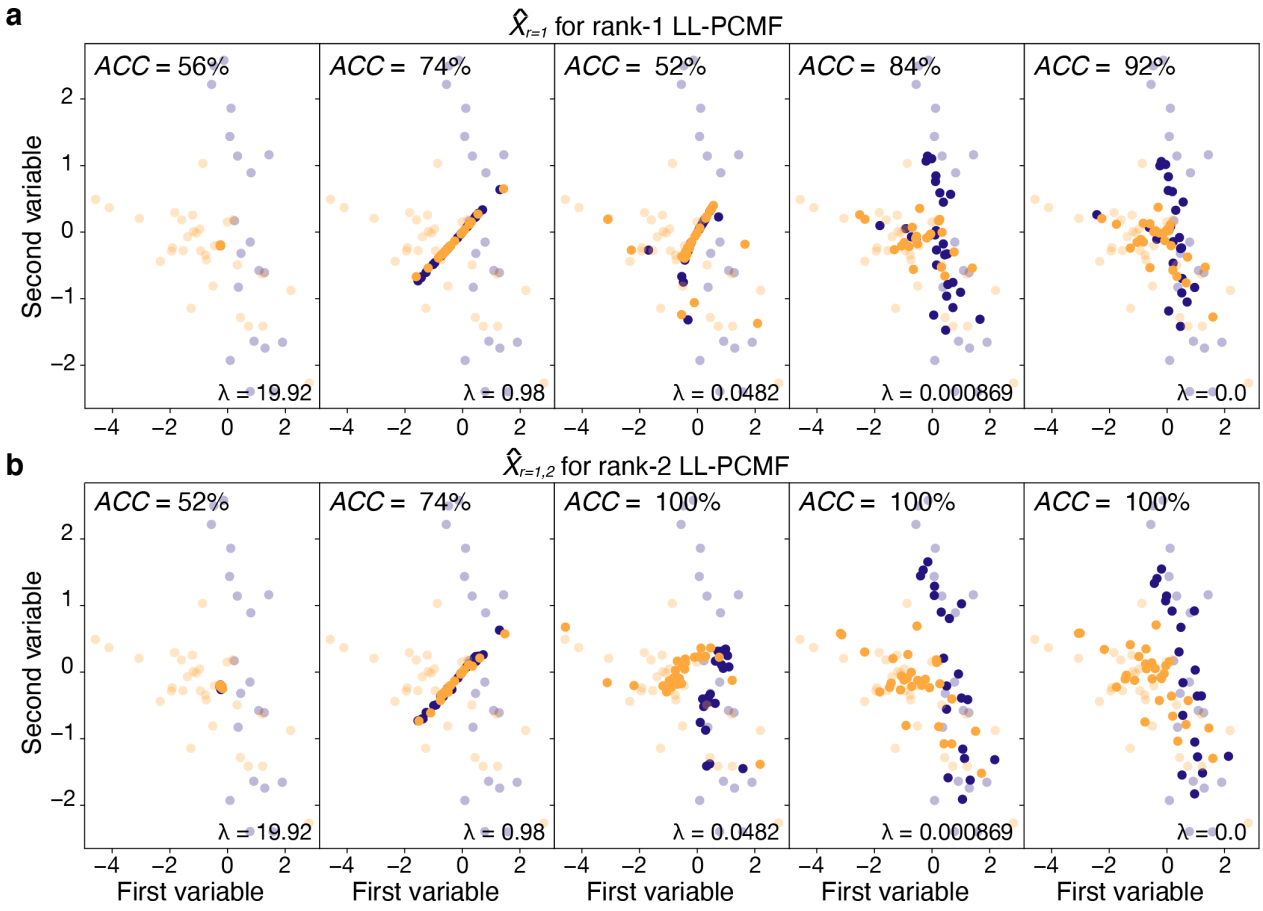

Figure 3: Example of rank-2 LL-PCMF using orthogonalized deflation (see Alg. 7) for a two-class problem. **a.** Reconstructed rank-1 LL-PCMF estimates, $\hat{X}_{r=1}$ (fit with $N.N. = 15; \rho = 1.5; \gamma = 2.0; K = 5$ ADMM iterations), evolving as λ decreases (five solutions along the path are shown). **b.** Reconstructed rank-2 LL-PCMF estimates, $\hat{X}_{r=1,2}$ (fit with $N.N. = 15; \rho = 1.5; \gamma = 2.0; K = 5$ ADMM iterations), evolving as λ decreases (five solutions along the path are shown). Dark dots indicate $\hat{X}_r$ colored by LL-PCMF predicted clusters. Light dots indicate X colored by true clusters. ACC indicates the accuracy of LL-PCMF predicted clusters at each λ path solution compared to the true cluster membership. Note rank-2 LL-PCMF shown in **b** achieves perfect cluster recover earlier in the solution path ($\lambda = 0.0482$) with more cluster separation compared to rank-1 LL-PCMF shown in **a**.

2.6 Convergence of PCMF, LL-PCMF, and P3CA Algorithms

We ran simulations for PCMF, LL-PCMF, and P3CA, and found that when fitting along a path of λs, the objective for these algorithms, although generally decreasing, can increase. We observe this most frequently near the first split point and, for PCMF, more for problem ranks that are smaller than warranted by the number of clusters (Appendix Fig. 5). We note that like many ADMM algorithms our approach seems to work well in practice, but a monotonic decrease of the objective is not guaranteed.

The data-generating parameters used for the datasets in Appendix Fig. 4a-b were 3 classes with $N_{\text{class}} = 50, \text{class} = 1, \ldots, 3$, $p = 20$, cluster centroids $\in \{-0.35, 0.0, 0.35\}$, and variable redundancy $\delta = 1$. In Appendix Fig. 4c, the data-generating parameters were 5 classes with $N_{\text{class}} = 20, \text{class} = 1, \ldots, 5$, $p = 50$, cluster centroids $\in \{-1.5, 1.5, 0.0, 0.01, -0.4\}$, and variable redundancy $\delta = 1$. For fitting PCMF, LL-PCMF, and P3CA, ρ was set to 1.0, γ was set to 2.0, the number of ADMM iterations per λ penalty was set to $K = 100$, and the number of nearest neighbors $N.N.$ was set to 25. For PCMF, we fit the model using problem ranks $1 \leq r \leq 10$ (Appendix Fig. 4a), and for LL-PCMF and P3CA, we fit the model using the $r = 1$ problem formulation (Appendix Fig. 4b). For the penalty path, we varied the value of λ along a path of 50 evenly spaced points in the interval $e^{[5,-20]}$ after initializing with $\lambda = \infty$ (10 iterations), such that the penalty decreased along the path of embedding solutions.

Note, in Appendix Fig. 4 we set $K = 100$ to ensure convergence at each λ, however, in practice a typical range for our implementation of the Algorithmic Regularization approach (see details in Appendix §2.7) is $K \in [5, 15]$. We use $K = 5$ for all other experiments shown in our figures and tables (as detailed in §5) unless specified otherwise in the specific legend or description.

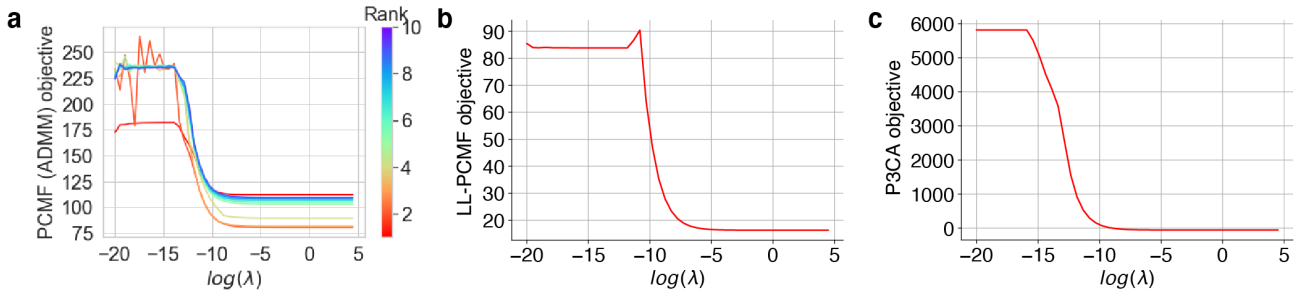

Figure 4: Empirical convergence of PCMF, LL-PCMF, and P3CA algorithms along λ path. **a.** PCMF fit to 3-class data ($p = 20$; $N_{\text{class}} = 50, \text{class} = 1, 2, 3$; $K = 100$; see Appendix §2.6 for further details) across a range of $1 \leq r \leq 10$. **b.** LL-PCMF fit to the same data using the same hyperparameters and $r = 1$. **c.** P3CA fit to 5-class multiview data ($p = 20$; $N_{\text{class}} = 50$, class $= 1, \ldots, 5$; $K = 100$; see text for further details) using the same hyperparameters and $r = 1$. In all cases while the objective trends downward, it can locally increase with increasing iteration and decreasing λ, empirically demonstrating that the path-wise objective is not strictly decreasing. K denotes the number of ADMM iterations per λ.

2.7 Algorithmic Regularization (AR) vs. Alternating Direction Method of Multipliers (ADMM)

A warm-started ADMM (alternating direction method of multipliers (Glowinski and Marroco, 1975, Gabay and Mercier, 1976, Boyd et al., 2011b) approach—Algorithmic Regularization—was recently introduced to enable feasible computation of dense convex clustering λ paths, speeding convergence more than 100-fold (Weylandt et al., 2020).

Note that in Main Text Algorithms 1–2 (reproduced here in Appendix Algorithms 1 and 6) and Appendix Algorithm 3 and 2, it is possible to obtain Algorithmic Regularization (AR) as presented previously (Weylandt et al., 2020) by setting $K = 1$, or to approach convergence at each value of λ by setting K large. We prefer to run $K = 5$ to $K = 15$ iterations at each value of λ, allowing some convergence at each value along the path. Appendix Fig. 5 compares AR ($K = 5$) paths to ADMM ($K = 100$) paths using Algorithm 1 (PCMF) to solve a four cluster problem; we show the first two variables out of $p = 200$ for data with 25 observations per class and means $\in \{-1.0, -0.4, 0.4, 1.0\}$, and compare the resulting number of clusters, the likelihood function (objective), the coefficient paths, and the estimated dendrograms for AR vs. ADMM.

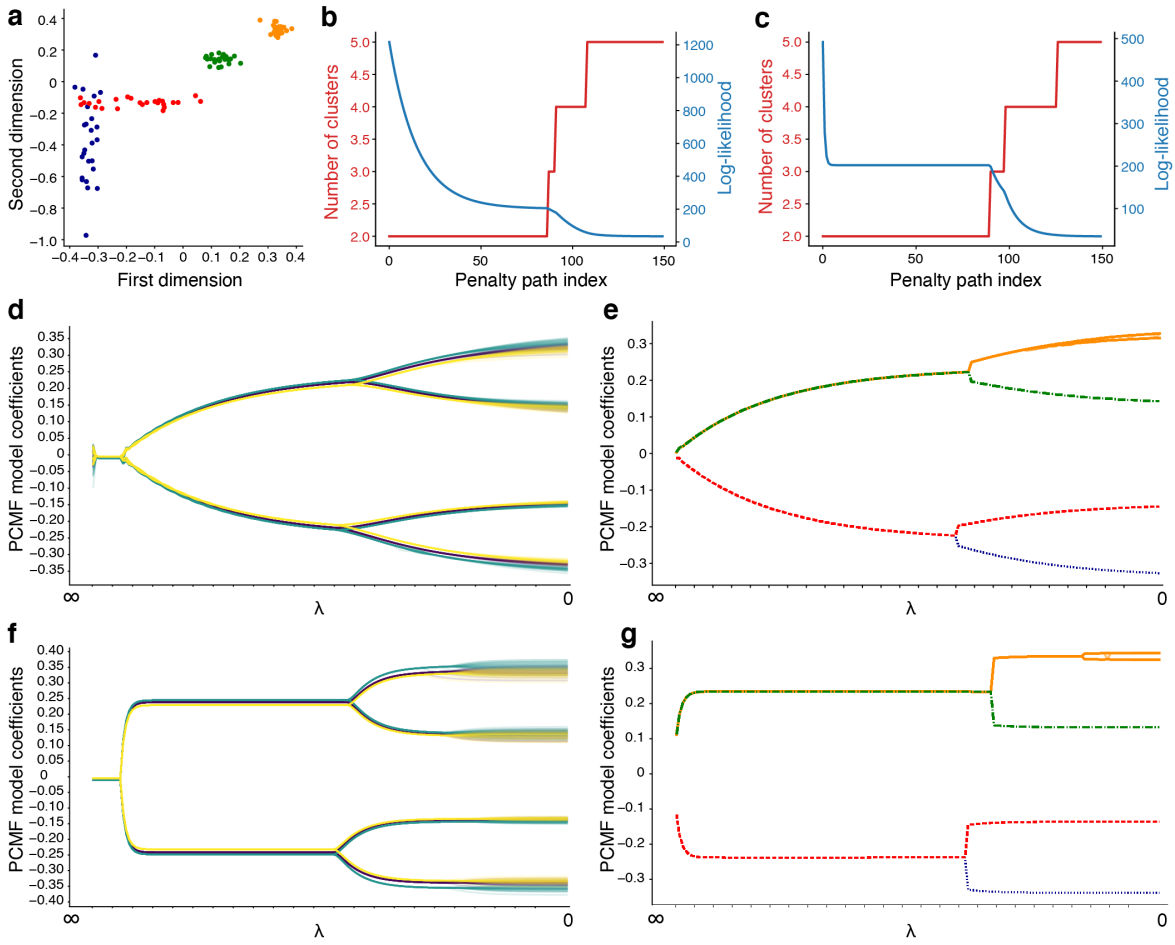

Figure 5: Comparison of PCMF using algorithmic regularization (AR) and PCMF using alternating direction method of multipliers (ADMM). **a.** Ground truth data for 4-class problem ($p = 200$; 25 observations per class; $r = 1$). **b–c.** Model selection for PCMF with **b.** AR or **c.** ADMM. Plots show the index along the penalty path versus the log-likelihood objective (blue) and the number of clusters (red) chosen using our correlation statistic model selection procedure. **d.** PCMF path (first three variables shown) and **e.** dendrogram (first variable shown) for PCMF using AR. **f.** PCMF path (first three variables shown) and **g.** dendrogram (first variable shown) for PCMF using ADMM. The resulting number of clusters chosen along the path, the evolution of the likelihood function (objective), the resulting coefficient paths, and the estimated dendrograms are qualitatively similar, with the AR estimates appearing as smoothed versions of the converged ADMM solutions.

2.8 Computational Complexity

Each **PCMF ADMM** iteration has a worst case complexity of $\mathcal{O}(N^2p)$ per iteration for $q \in \{1, 2\}$ and $\mathcal{O}(N^2p + log\ p)$ for $q = \infty$ due to the ADMM convex clustering subproblem (Min et al., 2018), and since the tSVD is $\mathcal{O}(N^2p)$. As the AR approach runs to convergence by warm-starting along the path of λ rather than iterating to convergence at each λ, for a path of length M this leads to worst case complexity of $\mathcal{O}(N^2pMK)$ for $q \in \{1, 2\}$ and $\mathcal{O}(N^2pMK + log\ p)$ for $q = \infty$ (for K iterations per λ, which is typically small, e.g., $K \in [5, 15]$). This worst-case can be improved upon significantly in practice (Chi and Lange, 2015) with the local weights and nearest neighbors approaches (rendering Cholesky factorization in the ADMM sparse), as well as by caching the Cholesky decomposition (Chi and Lange, 2015, Weylandt et al., 2020).

The **Penalized Alternating Least Squares (PALS)** formulation used for LL-PCMF relies on alternating solves of convex clustering problems using ADMM and projections onto constraint sets. The convex clustering steps dominate the complexity, again each with worst case complexity $\mathcal{O}(N^2p)$ per iteration for $q \in \{1, 2\}$ and $\mathcal{O}(N^2p + log\ p)$ for $q = \infty$ due to the ADMM convex clustering subproblem (Min et al., 2018), leading to the same

overall path-wise complexity as the PCMF ADMM approach: $\mathcal{O}(N^2pMK)$ for $q \in \{1, 2\}$ and $\mathcal{O}(N^2pMK + log\ p)$ for $q = \infty$. LL-P3CA has similar complexity: $\mathcal{O}(N^2p_xMK)$ for $q \in \{1, 2\}$ and $\mathcal{O}(N^2p_x + log\ p)$ for $q = \infty$ where p_x is the larger of the two datasets number of variables (that is, w.l.o.g. we let $p_x \geq p_y$).

This overall quadratic dependence on N in these approaches at first seems quite limiting if we would like to scale to problems with large N, and might lead one to question the decision to use ADMM approaches given the existence of potentially faster AMA (Chi and Lange, 2015) and semismooth Newton based augmented Lagrangian methods (Sun et al., 2021) for standard convex clustering (that is, without joint embedding). However, as shown in very recent work on scaling standard convex clustering (Fodor et al., 2022), there are quite significant advantages to the ADMM approach's ready scalability via consensus-optimization approaches. In our case, ADMM's established performance on non-convex problems and our need to additionally manage the $\mathcal{O}(N^2p)$ time complexity of embedding both provide significant additional justifications to using this approach.

Thus, to overcome the potentially limiting quadratic dependence on N due to convex clustering with ADMM and our embedding approach, we developed a **consensus ADMM** approach for our methods that allows us to run batch-wise updates in parallel on mini-batches with size N_b (with $N_b << N$). In this case, the ADMM consensus approach is dominated by the primal updates that it runs in parallel, allowing worst case $\mathcal{O}({N_b}^2pMK)$ for $q \in \{1, 2\}$ and $\mathcal{O}({N_b}^2pMK + log\ p)$ for $q = \infty$. Empirically, this means that our consensus ADMM method can run on cases with large N and p where state-of-the-art methods for standard convex clustering fail (e.g., CARP (Weylandt et al., 2020); see Main Text Table 3), and scale to problems larger than those shown to date for standard convex clustering (Sun et al., 2021) while—unlike any other existing convex clustering methods—also solving the joint embedding problem.

2.9 Pathwise Dendrogram Algorithm for Model Selection

Pathwise dendrogram algorithm:

Let $m = 1, \ldots, M$ index the decreasing path $\{\lambda\}$ such that $\{\lambda\} = \{\lambda_1 > \cdots > \lambda_m > \cdots > \lambda_M \geq 0\}$. Then to estimate a dendrogram from the smooth paths of the dual variables $\{G_{\lambda_m}\}$, we start at λ_1 (chosen large enough to yield only one cluster) and then proceed along the decreasing path of $\lambda_m > \lambda_{m+1}$. At each step, we make a binary choice between (a) keeping the same number of clusters $c_{m+1} \leftarrow c_m$, or (b) augmenting the number of clusters $c_{m+1} \leftarrow c_m + 1$. To make this choice, we find the partitions of the graph defined by $G_{\lambda_{m+1}}$ (see Chi and Lange (2015)) into c_m and c_{m+1} clusters, and then compare:

$$\text{loglik}_1(X, \widehat{X}(c_m), \lambda_m) = \frac{1}{2}\|X - \widehat{X}(c_m)\|_F^2 + \lambda_m \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\widehat{X}_{i\cdot}(c_m) - \widehat{X}_{j\cdot}(c_m)\|_q, \tag{43}$$

and

$$\text{loglik}_2(X, \widehat{X}(c_m + 1), \lambda_m) = \frac{1}{2}\|X - \widehat{X}(c_m + 1)\|_F^2 + \lambda_m \sum_{(i,j)\in\mathcal{E}} w_{ij}\|\widehat{X}_{i\cdot}(c_m + 1) - \widehat{X}_{j\cdot}(c_m + 1)\|_q, \tag{44}$$

where $\widehat{X}(c_m)$ is $\widehat{X}$ clustered so that its rows are replaced by c_m unique centroids (that is, $\widehat{X}$ is clustered to have exactly c_m unique rows). If $\text{loglik}_1 > \text{loglik}_2$ then $c_{m+1} \leftarrow c_m + 1$, otherwise $c_{m+1} \leftarrow c_m$. This ensures a dendrogram fit along the paths with knots in the number of clusters appearing only when the improvement in model fit exceeds the additional cost of adding another cluster to the penalty. In our experiments, to approximate the portion of the graph defined by the $\{G_{\lambda_m}\}$, we use spectral clustering (Appendix Fig. 5). As this algorithm uses K-means clustering on the eigenvectors of the affinity matrix estimated from the differences $G_{\lambda_m} = D\widehat{X}_{\lambda_m}$, it introduces random variation in the resulting dendrograms. We therefore choose to take the median of several runs of the pathwise dendrogram algorithm as the final estimate of the dendrogram, which we refer to as DENDROGRAM($\{G_\lambda\}$).

Although this approach performs well in our experiments, other approaches to graph partitioning applied to the graph defined by $\{G_{\lambda_m}\}$ are worth exploring, as they may provide more stable or more efficient approaches. Finally, it is critical to note that the dendrogram denotes the evolution of the centroids, not the individual observations (although these do become the individual observations in the limit $\lambda \to 0$. It is possible, although rare, for observations to switch class membership as the centroid dendrogram is estimated, thus while the result will always be a tree structure, we take the end-leaf membership as final assignment in these cases and trace membership back

up the estimated dendrogram post hoc for such cases to avoid ambiguity and satisfy the definition of a dendrogram.

A correlation-test-statistic-based heuristic for the number of clusters:

Previous work has shown a close relationship between convex clustering and single-linkage hierarchical clustering by examining the dual problems of the convex clustering optimization problem and a related problem that has the same connected component structure as single-linkage hierarchical clustering (see Lemmas 2-4 in Tan and Witten (2015)). Other work developing correlation tests of significance for the number of connected components in the graphical lasso (Friedman et al., 2007) fit along a path of penalty parameters that control model sparsity, have shown that this problem is also equivalent to thresholded single-linkage hierarchical clustering on correlations (G'Sell et al., 2013). Taken together, these findings suggest that extending the same correlation test statistic for the graphical lasso to the convex clustering problem in order to choose the best number of clusters for a given dataset may be fruitful.

In particular, let $\{G_{\lambda_t}\}$ for $t = 1, \ldots, T$ be the values or "knots" at which the number of clusters chosen by the pathwise dendrogram algorithm change along path $\{G_{\lambda_m}\}$ (so $G_{\lambda_m} \in \{G_{\lambda_t}\}$ if and only if $G_{\lambda_m} \to G_{\lambda_{m+1}} \implies c_{m+1} = c_m + 1$), then we note that as in (G'Sell et al., 2013) the $\lambda_1 > \cdots > \lambda_t > \cdots > \lambda_T$ correspond to the subset of knots at which the connected components of the estimate change. This naturally leads to a set of hypotheses, $H_1, \ldots, H_T$, where the hypothesis H_t is that each connected component of the true dendrogram is contained within the connected component defined by the estimated $\textsc{Dendrogram}(\{G_\lambda\})$ $\forall \lambda < \lambda_t$. To test these hypotheses, we note that the convex clustering problem is related to the group lasso estimator on the rows of G_λ, yielding the potential test statistic:

$$T_t = N\lambda_t(\lambda_t - \lambda_{t+1}), \tag{45}$$

an adaptation of the correlation test originally developed for the lasso in (Lockhart et al., 2014). However, as our problem is nonconvex due to the SVD constraint on convex clustering, and as we note that unlike the graphical lasso problem our observations can switch components along the path (hypotheses are not strictly nested), here we note this approach as an effective (see Appendix Tables 3 and 4 below) heuristic rather than an asymptotic result.

Model selection results for choosing the number of clusters:

We performed experiments evaluating the accuracy of our correlation-statistic-based model selection approach in synthetic datasets and four of our real-world datasets from Main Text Table 1 and compared it against three standard model selection criteria for five other clustering methods (including deep clustering).

We compare our model selection approach to three standard model selection criteria for clustering (taking the maximum Silhouette, Calinski-Harabasz, and Davies-Bouldin scores over a range of a possible 1-12 clusters) applied to two sequential embedding and clustering methods (PCA + K-means and PCA + Spectral clustering), to standard Spectral clustering, to subspace clustering (Elastic Subspace Clustering), and clustering with deep neural networks (DEC). These are generated to fall within the "non-trivial" clustering regime described in Definition 1 of Appendix §3. We found our approach compares very favorably with standard methods applied across the other methods on these challenging clustering problems (Appendix Table 3). We thus present this as an additional strength of our method: if model selection is desired, the method we propose appears quite robust and effective. Our model selection procedure is thus consistent with the clear visible separation of the coefficient paths and dendrograms into the correct number of clusters that we see throughout the Main Text and Appendix examples.

To further validate our model selection approach, we applied the same comparison to four real-world datasets used in the paper: NCI, SRBCT, MouseOrgans, and GBMBreastLung (Appendix Table 4). We find that our model selection approach, coupled with PCMF and LL-PCMF, yields model selection results that compare quite favorably to standard cluster selection metrics (maximum Silhouette, Calinski-Harabasz, and Davies-Bouldin scores taken over 2-12 clusters) applied to applied to two sequential embedding and clustering methods (PCA + K-means and PCA + Spectral clustering), to standard Spectral clustering, to subspace clustering (Elastic Subspace Clustering), and clustering with deep neural networks (DEC).

Table 3: Comparison of Model Selection Across Methods on Synthetic Data. Each value in the table is the mean ± standard deviation for 10 independently generated synthetic datasets of the specified number of clusters with and each cluster containing samples. Correlation Statistic is our model selection approach described in Appendix §2.9. C.-H., Calinski-Harabasz; Corr. Stat., Correlation Statistic, D.-B., Davies-Bouldin; Sil., silhouette

		2 clusters	3 clusters	4 clusters	5 clusters	6 clusters
PCMF	Corr. Stat.	2.0 ± 0.0	3.1 ± 0.3	4.0 ± 0.0	5.6 ± 0.6	6.0 ± 0.0
LL-PCMF	Corr. Stat.	2.5 ± 0.5	3.1 ± 0.3	3.8 ± 1.6	3.9 ± 2.1	5.4 ± 1.4
PCA + K-means	Sil.	8.2 ± 0.8	8.2 ± 1.8	9.9 ± 0.8	10.7 ± 0.6	10.9 ± 0.3
PCA + K-means	C.-H.	10.2 ± 0.98	11.0 ± 0.0	10.5 ± 0.50	6.5 ± 4.5	2.0 ± 0.0
PCA + K-means	D.-B.	2.0 ± 0.0	2.0 ± 0.0	2.4 ± 0.49	2.7 ± 0.8	3.9 ± 1.1
Spectral	Sil.	6.2 ± 1.6	6.3 ± 0.9	8.8 ± 1.3	9.8 ± 1.0	9.1 ± 2.6
Spectral	C.-H.	7.8 ± 0.40	10.8 ± 0.4	10.3 ± 1.0	9.0 ± 2.5	6.5 ± 3.5
Spectral	D.-B.	2.0 ± 0.0	2.4 ± 0.49	3.8 ± 0.7	3.3 ± 0.6	3.5 ± 1.3
PCA + Spectral	Sil.	5.3 ± 1.62	6.6 ± 1.50	8.5 ± 1.12	9.7 ± 1.0	11.0 ± 0.0
PCA + Spectral	C.-H.	8.2 ± 0.98	11.0 ± 0.0	9.9 ± 0.83	9.8 ± 2.6	3.0 ± 2.7
PCA + Spectral	D.-B.	2.0 ± 0.0	2.4 ± 0.49	3.1 ± 0.94	4.3 ± 0.9	5.0 ± 0.7
Elastic Subspace	Sil.	4.0 ± 0.0	5.0 ± 0.0	7.0 ± 0.0	9.0 ± 0.0	10.4 ± 0.5
Elastic Subspace	C.-H.	4.1 ± 0.3	5.0 ± 0.0	7.5 ± 0.9	9.0 ± 0.0	10.4 ± 0.5
Elastic Subspace	D.-B.	9.8 ± 0.98	2.8 ± 2.4	3.8 ± 2.4	2.0 ± 0.0	3.8 ± 1.4
DEC	Sil.	3.7 ± 2.69	6.6 ± 2.37	6.7 ± 2.32	7.4 ± 3.0	5.2 ± 2.7
DEC	C.-H.	4.4 ± 2.84	5.6 ± 1.90	7.4 ± 3.2	7.9 ± 2.1	5.2 ± 2.5
DEC	D.-B.	2.6 ± 0.8	7.9 ± 2.43	5.2 ± 2.2	5.7 ± 2.4	7.9 ± 2.6

Table 4: Model selection across methods on real-world datasets. Correlation Statistic (Corr. Stat.) is described in §2.9. C.-H., Calinski-Harabasz; D.-B., Davies-Bouldin; Sil., silhouette

		NCI	SRBCT	Mouse	Tumors
	Variables (p)	6,830	2,318	16,944	11,931
	Samples (N)	64	88	125	142
True Number of Classes		9	4	7	3
PCMF	Corr. Stat.	7	5	7	3
LL-PCMF	Corr. Stat.	4	4	8	3
PCA + K-means	Sil.	11	10	2	11
PCA + K-means	C.-H.	4	2	2	4
PCA + K-means	D.-B.	6	2	2	6
Spectral	Sil.	5	7	2	5
Spectral	C.-H.	3	2	2	3
Spectral	D.-B.	3	2	2	3
PCA + Spectral	Sil.	2	7	2	2
PCA + Spectral	C.-H.	2	5	2	2
PCA + Spectral	D.-B.	3	2	2	3
Elastic Subspace	Sil.	11	7	2	11
Elastic Subspace	C.-H.	2	7	3	2
Elastic Subspace	D.-B.	3	2	2	3
DEC	Sil.	10	11	3	7
DEC	C.-H.	5	11	3	7
DEC	D.-B.	9	10	2	6

Model selection for known number of clusters:

Given a prespecified number of clusters c (if for example, we know the number of desired clusters beforehand), we may want to choose a best model conditional on c. As there may be more than one value of m for which $c = c_m$, we choose the m that solves

$$\underset{m}{\text{minimize}} \ \text{loglik}_1 \left(X, \widehat{X}(c_m), \lambda_m \right) \ \text{subject to } c_m = c. \tag{46}$$

3 THEORETICAL PROOF: PCMF DOMINATES CONVEX CLUSTERING FOR P>N (LDL) DATA

Here we offer a constructive proof for inadmissibility of convex clustering in the case of asymptotically "nontrivial" $p > N$ GMM clustering by showing that PCMF dominates convex clustering for such data using results from random matrix theory (RMT) in the large dimensional limit (LDL) regime. This regime (also known as the Kolmogorov regime or the $p > N$ regime) corresponds to the case where the number of variables p are on the same order as the number of observations N, specifically to the asymptotic case $p/N \to c \in (0, \infty)$ as $N, p \to \infty$. Note that $\|\cdot\|$ denotes the operator norm and that we have overloaded $O(\cdot)$ here to indicate $O(\cdot)$, $\Theta(\cdot)$, or $\Omega(\cdot)$ from standard computer science notation (its exact meaning will be clear from context).

Definition 1 "Nontrivial" GMM data (Couillet and Liao, 2022). We consider observing N i.i.d. data vectors $\mathbf{x}_i \in \mathbb{R}^p$ drawn from the K-class GMM with fixed class sizes $N_1, \ldots, N_K$ (with $\sum_{k=1}^K N_k = N$) gathered in data matrix $X = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T \in \mathbb{R}^{N \times p}$, with $p \sim N$ or $p > N$ such that $p/N \to c \in (0, \infty)$ and $N_a/N \to c_a \in (0, 1)$ as $N, N_a, p \to \infty$. Letting $\mathcal{C}_a$ be the set of observations from class a for $a \in \{1, \ldots, K\}$ such that $\mathbf{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_a, C_a) \iff \mathbf{x}_i \in \mathcal{C}_a$ with $C_1, \ldots, C_k$ distinct and of bounded norm. To ensure that cluster separation is nontrivial as $N, p \to \infty$ we take $\|\mathbf{x}_i\|$ to be of order $O(p^{1/2})$ and $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\| = O(1)$ for $a, b \in 1, \ldots, K; a \neq b$. Note that Definition 1 deviates somewhat here from the typical, probabilistic GMM definition to highlight that we are drawing from different clusters. Also note that for lightness of notation $\mathbf{x}_i$ denotes a row vector of X (that is, $\mathbf{x}_i = X_{i\cdot}$). Note that this case is considered asymtotically "nontrivial" because it ensures that asymptotic clustering is neither trivially easy nor impossible when $p, N \to \infty$, as the cases below will demonstrate.

Proposition 1. For clustering the nontrivial Gaussian Mixture Model (GMM) data in the low dimensional limit (LDL) regime, PCMF asymptotically dominates standard convex clustering.

Proof. Our proof proceeds in two parts, closely following recent results in RMT (see Ch. 2 and Ch. 4 in Couillet and Liao (2022)). We first show that convex clustering asymptotically fails to discriminate classes in the nontrivial GMM clustering problem, and then second show that PCMF asymtotically succeeds for the exact same regime. Together these imply that PCMF dominates convex clustering as an estimator for nontrivially clustered GMM data in the LDL regime.

We first note that the GMM assumption in many cases can be taken to be equivalent to requiring $\mathbf{x}_i \in \mathcal{C}_a$: $\mathbf{x}_i = \boldsymbol{\mu}_a + C_a^{1/2}\mathbf{z}_i$ where $\mathbf{z}_i$ is a random vector with i.i.d. zero mean, unit variance, and suitably bounded higher-order moment entries. Intriguingly, recent RMT results show that in the LDL regime this model has the same asymptotic statistics as a number of much more complicated models (see, e.g. Ch. 8 in Couillet and Liao (2022)). Next we define $C^\circ = \sum_{a=1}^k \frac{N_a}{N} C_a$ and $C_a^\circ = C_a - C^\circ$ as the average and centered covariances, and $\boldsymbol{\psi}$ as a vector with elements $[\boldsymbol{\psi}]_i = \mathbf{z}_i^T C_a \mathbf{z}_i - \operatorname{tr} C_a/p$, and note that by straightforward central limit theorem arguments $[\boldsymbol{\psi}]_i = O(p^{-1/2})$.

With these definitions in place, we again consider the standard convex clustering optimization problem, re-expressed in terms of the entries and rows of X and $\widehat{X}$ as:

$$\underset{\widehat{X}}{\text{minimize}} \; \frac{1}{2} \sum_i \sum_j \|x_{ij} - \widehat{x}_{ij}\|_2^2 + \lambda \sum_{i<j} f\left(\frac{1}{p}\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\|_2^2\right). \tag{47}$$

Note we have subsumed the important weights w_{ij} and an element-wise square root into function $f(\cdot)$ and we normalize by the number of variables p (without loss of generality) to be in compliance with the defined normed quantities in Definition 1. For $i \neq j$ we can expand the terms inside $f(\cdot)$ "entry-wise" as:

$$\begin{aligned}
\frac{1}{p}\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\|_2^2 &= \frac{2}{p}\operatorname{tr} C^\circ + \frac{1}{p}\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2 + \frac{1}{p}\operatorname{tr}\,(C_a^\circ + C_b^\circ) - \frac{2}{p}\mathbf{z}_i^T C_a^{-1/2} C_b^{-1/2}\mathbf{z}_j + \frac{1}{p}\mathbf{z}_i^T C_a \mathbf{z}_i \\
&\quad - \frac{2}{p}\operatorname{tr} C_a + \frac{1}{2}\mathbf{z}_j^T C_b \mathbf{z}_j - \frac{1}{p}\operatorname{tr} C_b + \frac{2}{p}(\boldsymbol{\mu}_a - \boldsymbol{\mu}_b)^T (C_a^{1/2}\mathbf{z}_i - C_b^{1/2}\mathbf{z}_j) \\
&= \frac{2}{p} C^\circ + O(p^{-1/2}).
\end{aligned} \tag{48}$$

The final equality follows by considering each term given the nontrivial GMM data assumptions and RMT data representation made above. Note that critically this means if we consider the entry-wise distances in the

$p, N \to \infty; p/N \to c$ regime, all entries are dominated by the constant $2\text{tr } C^\circ/p$, which is $O(1)$, regardless of the values that a and b take, and $\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2/p = O(p^{-1})$ which is dominated by the noise terms which are $O(p^{1/2})$.

This shows that an entry-wise distance approach like that taken in convex clustering does not work for discrimination in this regime.

However, if we consider information "spread across" the many variables in their spectrum, we see that there is important discriminative information available in the low-rank structure of the Euclidean distances. In particular, letting $M = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k] \in \mathbb{R}^{p\times k}$, $\mathbf{d} = \text{diag}(JM^TZ)$, $W = [C_1^{1/2}Z_1, \dots, C_k^{1/2}Z_k] \in \mathbb{R}^{p\times N}$, $\mathbf{t} = \{\text{tr } C_a^\circ/p\}_{a=1}^k \in \mathbb{R}^k$, and defining $A - \text{diag}(\cdot)$ as an operator that returns matrix A with diagonal entries set to zero, we can look "matrix-wise":

$$\begin{aligned}\left\{\frac{1}{p}\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\|_2^2\right\}_{i,j=1}^N &= \frac{2}{p}\text{tr } C^\circ \cdot \mathbf{1}_N\mathbf{1}_N^T + \frac{1}{p}J\{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2\}_{a,b=1}^k J^T \\ &\quad + (\boldsymbol{\psi} + J\mathbf{t})\mathbf{1}_N^T + \mathbf{1}_N(\boldsymbol{\psi} + J\mathbf{t})^T - \frac{2}{p}W^TW \\ &\quad + \frac{2}{p}(\mathbf{d}\mathbf{1}_N^T + \mathbf{1}_N\mathbf{d}^T) - \frac{2}{p}(JM^TW + W^TMJ) - \text{diag}(\cdot). \end{aligned} \tag{49}$$

In this case, although the matrix is again dominated by the $O(N)$-norm matrix $2\text{tr } C^\circ/p \cdot \mathbf{1}_N\mathbf{1}_N^T$, there is now critically usable information in the $O(N^{-1/2})$-norm rank-2 matrix $(\boldsymbol{\psi} + J\mathbf{t})\mathbf{1}_N^T + \mathbf{1}_N(\boldsymbol{\psi} + J\mathbf{t})^T$. Indeed, it is easily shown this matrix gives access to the covariance traces through terms that are $O(p^{1/2})$ by looking at the second dominant eigenvector of the Euclidean matrix (as commonly done in spectral clustering). Further, the smaller order $O(1)$ terms:

$$\frac{1}{p}J\{\|\boldsymbol{\mu}_a - \boldsymbol{\mu}_b\|_2^2\}_{a,b=1}^k J^T + \frac{2}{p}(\mathbf{d}\mathbf{1}_N^T + \mathbf{1}_N\mathbf{d}^T) - \frac{2}{p}(JM^TW + W^TMJ) - \text{diag}(\cdot), \tag{50}$$

contain usable and asymptotically available discriminative information about the means (Couillet and Liao, 2022).

PCMF—which relies on using a rank-r embedding $\widehat{X} \in \mathcal{M}_r$—gives direct access to this "matrix-wise" spectral information (through the tSVD), and therefore asymptotically allows discrimination. PCMF thus asymptotically dominates "element-wise" clustering methods like convex clustering in the LDL regime, as the latter cannot discriminate clusters in the nontrivial GMM clustering problem. □

Proposition 2. For clustering the nontrivial GMM data in the LDL regime, the locally linear relaxation (LL-PCMF) asymptotically dominates convex clustering.

Proof. The proof follows readily from showing that the spectral information available in PCMF is also available to the LL-PCMF embedding (e.g., that LL-PCMF recovers PCMF as a special case). First note that the deflation and augmentation scheme used in LL-PCMF allows each u_i and $\mathbf{v}_i$ to freely approximate the data, constrained only their relationships to other observations through the kernel induced by the local weights of the convex clustering (or relaxed convex clustering) penalty (see, e.g. (27)). This induces the following locally linear weighting scheme:

$$\underset{\widehat{X}}{\text{minimize}}\ \frac{1}{2}\sum_i\sum_j \|x_{ij} - w_{ij}\widehat{x}_{ij}\|_2^2 + \lambda\sum_{i<j} f\left(\frac{1}{p}\|\widehat{\mathbf{x}}_i - \widehat{\mathbf{x}}_j\|_2^2\right). \tag{51}$$

In the standard convex clustering formulation where $\widehat{X}$ is unconstrained, this results in the previously described clustering with local fitting behavior (Hocking et al., 2011). Considering the LL-PCMF formulation without the convex penalty relaxation, it is easy to see that by setting all weights $w_{ij} = 1$ for all i, j we recover a sequential rank-1 deflation scheme (Mackey, 2008, Witten et al., 2009) to solving the uniformly-weighted PCMF problem, and thus our method inherits the results of Proposition 1. Finally, to accommodate the relaxed LL-PCMF formulation, we note that the non-relaxed and relaxed penalties differ substantively only in the multiplicative cross-terms of the penalty (terms that contain $u_iu_j\mathbf{v}_i^T\mathbf{v}_j, i, j = 1, \dots, N$ in the non-relaxed case) being relaxed to additive terms (terms that instead contain $u_iu_j - \mathbf{v}_i^T\mathbf{v}_j, i, j = 1, \dots, N$). This strictly expands the solution space for the u_i and $\mathbf{v}_i$ variables such that the relaxed LL-PCMF problem solutions contain the non-relaxed LL-PCMF solutions (of which the PCMF solutions are a subset as described above). This implies that LL-PCMF, by relying on a (local) low-rank embedding $\widehat{X} \in \mathcal{M}$ that gives "matrix-wise" access to the (local) spectral information of the data and can asymptotically allow discrimination in the LDL regime given appropriately chosen weights. □

Proposition 3. PCMF generalizes kernel spectral clustering (kSC) to joint clustering and embedding. *Proof.* First we remind the reader that in PCMF the final hard clustering at each λ is performed on the weighted affinity matrix generated from the differences matrix $G = D\widehat{X}$ (defined by the dual variables as described in previous work (Chi and Lange, 2015). After thresholding, this procedure estimates the connected components of the resulting graph at each value of λ, specifically by looking at the low-rank spectrum of the Laplacian and clustering on it. Each hard clustering along the PCMF coefficient paths is thus equivalent to spectral clustering (or kernel spectral clustering for kernel or nearest neighbor weights w_{ij}). Indeed, in practice we have used an off-the-shelf kernel spectral clustering algorithm applied to the adjacency matrix of the thresholded affinity matrix for each λ. Thus, in the case $\lambda = 0$ where the solution exactly interpolates the data X (and so $G = DX$), PCMF recovers kernel spectral clustering on X exactly (with the kernel defined by the weights w_{ij} applied to data row differences $G = DX$ to generate the weighted affinity graph). In this sense, PCMF is trivially a generalization of kSC. □

Beyond this special case of $\lambda = 0$ being exactly kSC, there is some additional potential depth and intuition to the generalization. In particular, recall that spectral clustering generally proceeds by applying K-means clustering to the first r eigenvectors of the normalized adjacency matrix (or the last r eigenvectors of the normalized Laplacian). And recall that the convex clustering penalty with $q = 0$ has been shown to be equivalent to K-means clustering (up to an additional penalty term) (Tan and Witten, 2015), and convex clustering can thus be thought of as a form of "convex K-means". If we inspect the terms of the PCMF problem,

$$\underset{\widehat{X} \in \mathcal{M}_r}{\text{minimize}} \underbrace{\|(X - \widehat{X}\|_F^2}_{\text{tSVD}} + \lambda \sum_{i,j \in \mathcal{E}} \overbrace{w_{ij}}^{\text{kernel}} \underbrace{\|\widehat{X}_{i\cdot} - \widehat{X}_{j\cdot}\|_q}_{\text{convex k-means}}, \tag{52}$$

we see that the first term can be conceptualized as finding the rank-r spectrum (tSVD) of the data at the same time that the second term applies convex K-means clustering to the weighted adjacency matrix of this data projected onto its first r eigenvectors. This gives us some intuition for results like those shown in the Main Text Fig. 1, where PCMF clearly outperforms kSC on cluster-aware embedding by jointly learning a good embedding and appropriate clustering.

Finally, it is worth noting that problems like that detailed in the proof of Proposition 1 are precisely the approach that spectral clustering is meant to solve, as it uses spectral information spread out across the Euclidean distance matrix. However, we see here that by jointly shrinking towards the embedding and the clustering, we can potentially improve both while generalizing spectral clustering and not requiring that the number of clusters be preordained. Further, note that in standard spectral clustering, where $f(\cdot)$ is linear, if the means and covariance traces are equal (as in the synthetic "overlapping cluster" examples we show in our synthetic examples above), no spectral information can be retrieved that is of use in discrimination (as only mean and covariance trace information are available). This shows that a method where f is nonlinear (or contains nonlinear, local weights) is required for discrimination in such problems (and is consistent with the significant performance improvement afforded PCMF by using the nonlinear nearest neighbors approach to penalty weights).

Together, these propositions show how PCMF outperforms convex clustering asymptotically for a class of nontrivial $p > N$ problems and relates it formally and intuitively to kernel spectral clustering.

4 EXTENDED RESULTS

4.1 PCMF on Synthetic Data

PCMF, LL-PCMF, and P3CA: What are the hyperparameters and what is their impact?

The hyperparameters for the **globally linear PCMF (PCMF)** are augmented Lagrangian parameter (ρ), weight (γ), $N.N.$ (nearest neighbors), the number of ADMM iterations (K), and the convex clustering penalty (λ) and the global problem rank of the full low-rank embedding (r).

The hyperparameters for the **locally linear PCMF (LL-PCMF)** are augmented Lagrangian parameter (ρ), weight (γ), $N.N.$ (nearest neighbors), the number of ADMM iterations (K), and the convex clustering penalty (λ).

The hyperparameters for **pathwise canonical correlation analysis (P3CA)**, which we implement as a multiview generalization of LL-PCMF) are augmented Lagrangian parameter (ρ), weight (γ), $N.N.$ (nearest neighbors), the number of ADMM iterations (K), the number of CCA iterations (κ), and the convex clustering penalty (λ).

ρ is the augmented Lagrangian parameter. Although ADMM can in some cases be quite sensitive to choices of the augmented Lagrangian parameter ρ, we find our algorithm to be stable across a range of common ρ values (Appendix Fig. 7).

γ is the weight on each connection in the convex clustering penalty and influences the cluster fusion along the convex clustering path. In practice, we set this to $\gamma = 2.0$. Though this parameter could be tuned to optimize performance, we find the solution path to not be highly sensitive to small variations in the γ parameter (Appendix Fig. 6).

$N.N.$ is the nearest neighbors parameter and controls the number of connections and nested structure of the solution path; it influences cluster fusion and the time to solve the algorithm (Appendix Fig. 6).

K, the number of ADMM iterations per λ, controls convergence at each penalty on the path since we use Algorithmic Regularization (see Appendix §2.7; Appendix Fig. 5).

κ in P3CA is the number of CCA iterations per K, and influences convergence per ADMM iteration, K, to the CCA optimization. We set this to $\kappa = 2$.

λ, the convex clustering penalty controls the convergence to the clustering solution, as with algorithmic regularization (see Appendix §2.7) K is set to a smaller number than total convergence. Unless otherwise specified, we set the convex clustering penalty path to initialization using $K = 10$ $\lambda = \infty$ followed by 150 evenly spaced points in the interval $\lambda = e^{[10,-10]}$, such that the penalty decreased along the path of embedding solutions.

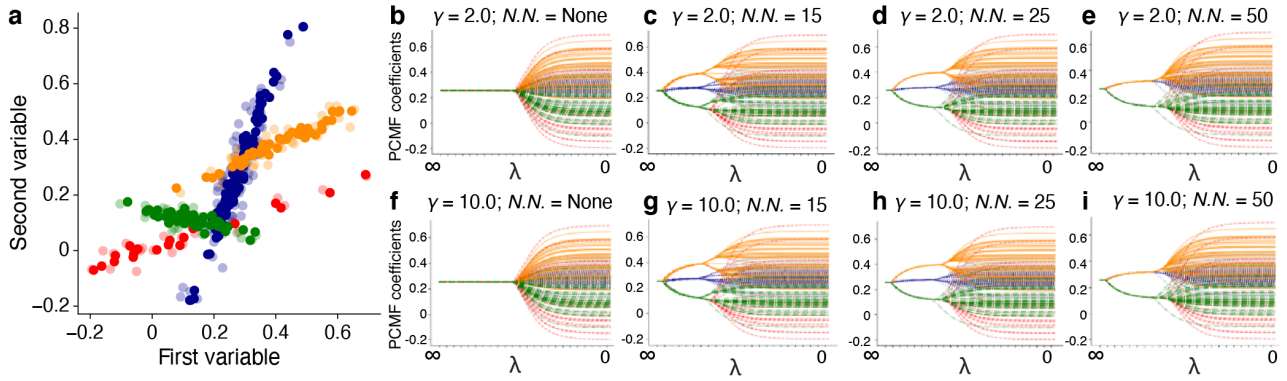

Figure 6: LL-PCMF varying nearest neighbors ($N.N.$) and weights (γ). **a.** Ground truth data for 4-class problem; $p = 20$; $N_1 = 100$ (blue), $N_2 = 50$ (orange), $N_3 = 20$ (red), $N_4 = 50$ (green). The first two variables of the data are plotted with light points showing raw data and dark points showing data reconstructed from low rank estimates, colored by true cluster membership (PCA rank $r = 5$). **b–i.** PCMF paths for variable 1 fit along decreasing penalty path ($\lambda = \infty$ to $\lambda = 0$) varying $N.N.$ and γ.

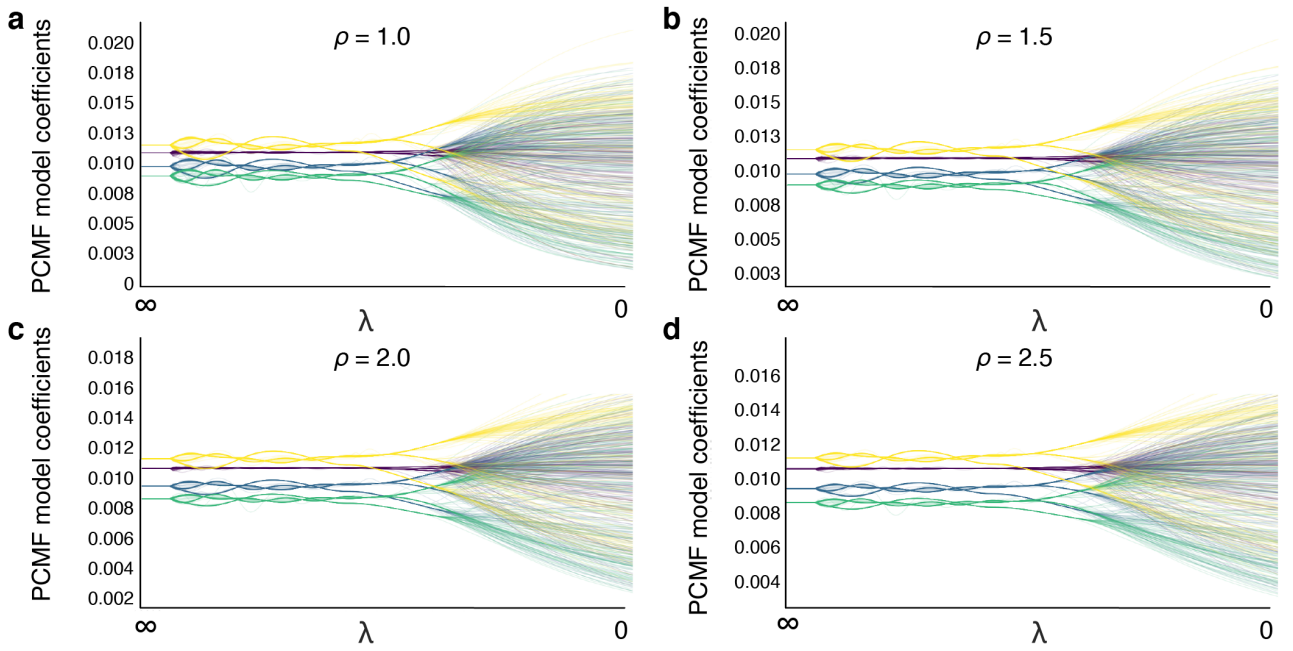

Figure 7: LL-PCMF model estimates for the Tumors dataset for four different augmented Lagrangian parameters ($\rho = 1.0, 1.5, 2.0, 2.5$) along the convex penalty (λ) path. Paths are stable across the four ρ values.

4.2 P3CA on Palmer Penguin dataset

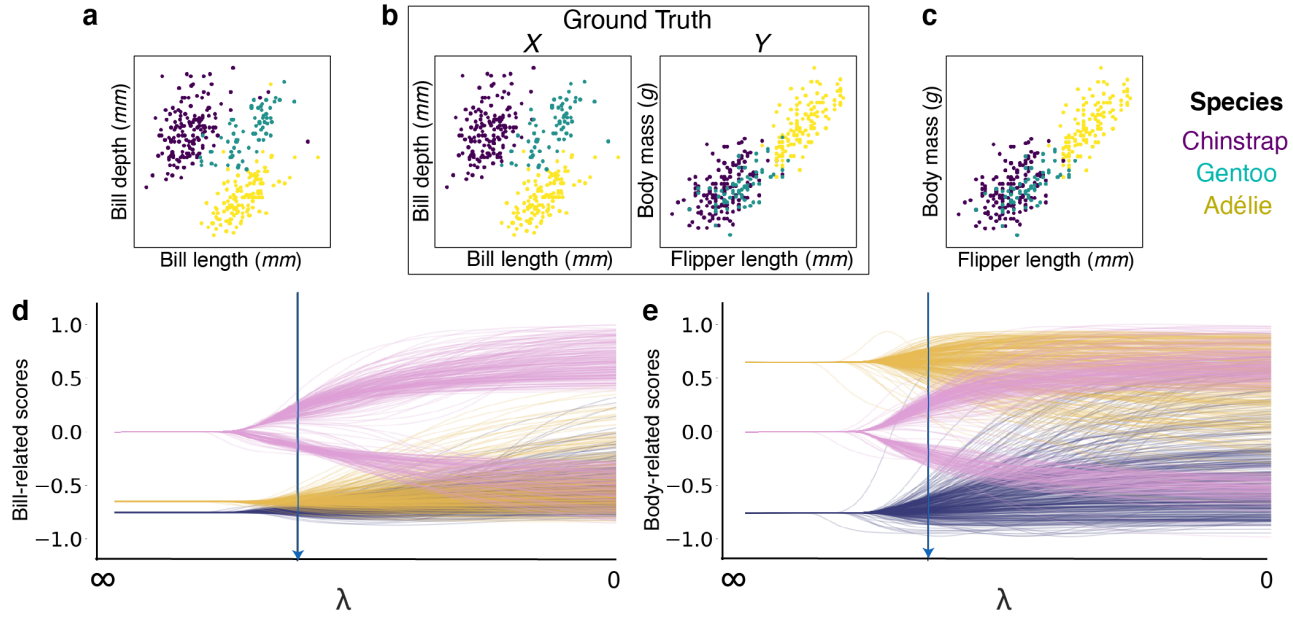

Figure 8: P3CA identifies penguin species-specific embeddings (left: bill-related; right: body-related). **a.** X data and **c.** Y data colored by P3CA clusters from λ path penalty (blue arrows). **b.** Ground truth clusters for X and Y. **d-e.** P3CA paths for X and Y (color indicates variable; pink is intercept).

Here we demonstrate an application of P3CA on a simple real world dataset, the Palmer Penguins dataset (Horst et al., 2020). The Palmer Penguin dataset consists of four body measurements in $N = 342$ penguins from three species (adélie, chinstrap, and gentoo penguins). For illustration of P3CA, we split the four measurements into two data views: X with bill length and bill depth ("bill-related" measurements) and Y with flipper length and body mass ("body-related" measurements) (Appendix Fig. 8b). We applied P3CA to this multiview phenotypic

dataset and found that P3CA recovered the three penguin species with high accuracy (accuracy = 98.25%; Main Text Table 2; Appendix Fig. 8a,c) and identified a hierarchy of cluster-specific embeddings ("Bill-related" in Appendix Fig. 8d and "Body-related" in Appendix Fig. 8e).

4.3 P3CA on Autism Spectrum Disorder (ASD) Neuroimaging Dataset

We highlight three cluster solutions along the P3CA path for behavior (as λ decreases) and show where the behavior-related P3CA variate (U) intercept splits—indicating at least two clusters (Appendix Fig. 9). We find the solution from sequential CCA+K-means (Appendix Fig. 9c) identifies a similar ASD-related brain-behavior embedding, but fails to fully separate the two clusters along this embedding. (similar to prior sequential CCA followed by clustering approaches in neuropsychiatry (Drysdale et al., 2017, Grosenick et al., 2019, Buch et al., 2023).

Further, the resulting correlations between ASD behaviors and the P3CA variate are significantly different across clusters (Appendix Table 5) as well as between ASD brain connections and the P3CA variate (Appendix Table 6)—consistent with known ASD subpopulation differences on RRBs and verbal IQ and prefrontal cortex to somatosensory cortex, posterior parietal cortex, and middle temporal gyrus. There are no ground truth clusters (biological subtypes of ASD is an open problem), so we measure cluster composition stability and U and V coefficients by randomly holding out 30% of data in 10 replicates (yielding 10 datasets of $N = 209$ X-Y) and calculating the P3CA path over 50 λ values in each subsample. For each subsample, we compare cluster assignment and subject-level U and V coefficients with those of the corresponding subjects when P3CA was calculated using the full dataset ($N = 299$, Appendix Fig. 9). We find that subject-level U and V coefficients are stable (cosine similarity of 0.93 ± 0.05 for U estimates and 0.97 ± 0.03 for V estimates).

Table 5: Correlations Between ASD Behaviors and P3CA variate. Bold text indicates correlations, $|r| \geq 0.20$. RRB, repetitive/restricted behaviors/interests

	Cluster 1	Cluster 2	Combined
RRB	**-0.50**	-0.05	**-0.45**
Verbal IQ	**-0.25**	**-0.63**	**-0.63**
Social Affect	-0.16	**-0.35**	**-0.58**

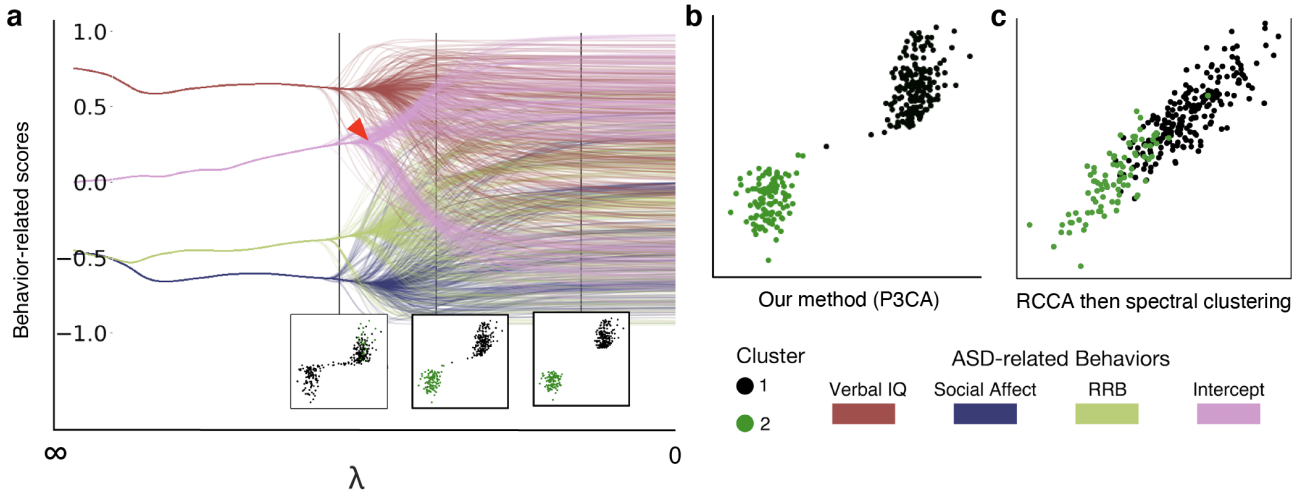

Figure 9: P3CA identifies distinct biological subtypes of ASD patients using multivariate clinical behavior ($X = 3$) features and functional brain connectivity ($Y = 20$ top features). **a.** P3CA path diagram of behavior-related P3CA scores with clustering solutions at λs indicated by line (see intercept splitting at red arrow). **b.** P3CA clustering solution separates ASD patients along the brain-behavior embedding. **c.** P3CA clusters are much better separated and more robust than sequential CCA followed by K-means clustering. Abbreviations: CCA, canonical correlation analysis; RRB, repetitive/restricted behaviors/interests

Table 6: Correlations between functional brain connectivity and P3CA variate. Bold text indicates correlations, $|r| \geq 0.20$. ACC, anterior cingulate cortex; antPFC, anterior prefrontal cortex; IFGorb, Orbital part of inferior frontal gyrus; IPL, inferior parietal lobe; ITG, inferior temporal gyrus; L, left; M1, primary motor cortex; MCC, midcingulate cortex; MFG, middle frontal gyrus; mOFC, medial orbitofrontal cortex; MOG, medial orbital gyrus; MTG, middle temporal gyrus; NAcc, nucleus accumbens; PCC, posterior cingulate cortex; PPC, posterior parietal cortex; R, right; S1, primary somatosensory cortex; SFG, superior frontal gyrus; SMA, supplementary motor area; SOG, superior orbital gyrus; SPL, superior parietal lobule; VLPFC, ventrolateral prefrontal cortex; VMPFC, ventromedial prefrontal cortex.

	Cluster 1	Cluster 2	Combined
L MTG – L thalamus	-0.16	-0.03	**-0.23**
L paracentral/S1 – R VMPFC/IFGorb	-0.04	**-0.20**	**-0.21**
R SOG – L cerebellum	-0.11	-0.05	-0.19
R ITG/MTG – R PCC	**-0.21**	-0.06	**-0.27**
L insula – L MOG	-0.13	-0.10	-0.05
R MTG/ITG – L MTG	-0.09	-0.14	**-0.21**
L antPFC/SFG – L NAcc	-0.03	-0.01	-0.19
L precentral/M1 – R VMPFC/mOFC	-0.09	-0.14	**-0.21**
R ACC/MCC – R VLPFC/MFG	-0.09	-0.05	-0.13
R VMPFC/IFGorb – R PPC/SPL	-0.18	**-0.25**	**-0.31**
R SMA/ACC – L MOG	-0.08	-0.14	-0.01
R VMPFC/IFGorb – L PPC/IPL	-0.11	**-0.39**	**-0.24**
R temp pole – L MTG	-0.04	-0.06	-0.15
R temp pole – L thalamus	**-0.28**	-0.07	**-0.21**
R VLPFC/MFG – R MTG	**-0.20**	-0.05	**-0.20**
R ACC/MCC – L ventral putamen	-0.00	-0.03	-0.02
L lingual – L VLPFC/IFGorb	-0.10	-0.04	-0.11
R VMPFC/IFGorb – L lingual	-0.14	-0.10	**-0.21**
L cuneus – L fusiform	-0.14	-0.10	**-0.24**
L lingual – L ACC	-0.03	-0.04	-0.03

5 EXPERIMENTAL METHODS AND DATASETS

Here we include details on experimental methods, synthetic dataset generation, and real-world datasets.

5.1 Synthetic Dataset Generation

We use 20 synthetic dataset types in our study.

Synthetic dataset 1. Main Text Figure 1: We generate synthetic data for a 3-class problem with $p = 20$ and $N_1 = 100$ (blue), $N_2 = 25$ (red), $N_3 = 25$ (orange), colored by true cluster membership in Main Text Fig. 1 with $\delta = 1.0$, $\sigma = 0.08$, and cluster centroids 1,2,3 $\in \{-0.35, 0.0, 0.35\}$ accordingly. The seed for data generation was set to ensure the three-class dataset is reproducible.

Synthetic datasets 2–13. Appendix Figure 1 and Appendix Table 1 include 12 types of two clustered-data in a single view with $p = 200$, or $p = 2000$ variables and variable redundancy set by $\delta = 0.5$ or $\delta = 0.2$, the fraction of variables containing signal. Varying p and δ, we generate data from two separate distributions with different slopes and cluster means in three variations: non-overlapping clusters (cluster centroids $\in \{-0.2, 0.2\}$; $N_1 = 50; N_2 = 50$), overlapping clusters (cluster centroids $\in \{-0.05, 0.05\}$; $N_1 = 50; N_2 = 50$), or non-overlapping but unbalanced cluster size (cluster centroids $\in \{-0.2, 0.2\}$; $N_1 = 80; N_2 = 20$).

For each dataset generated, the seed was set to ensure the two-cluster data was reproducible. Within a single dataset, for each cluster a random matrix of N x p with the specified mean and $\sigma = 0.075$ was generated. Next, $\mathbf{u}$ was generated as a random matrix of N x 1 and $\mathbf{v}$ was generated as a random matrix of p x 1. We selected

$\delta * p$ features from a randomly permuted order of the total p features and added $\mathbf{v}[i] * \mathbf{u}$ to $\mathbf{X}[:, i]$ where i is the feature (column). Finally, we standardized features over samples (rows) by removing the mean and scaling to unit variance, and added a column of ones as an intercept to the features. We compare results based on the adjusted rand index (ARI) and the normalized mutual information (NMI).

Synthetic datasets 14–15. Appendix Figures 2,5 and Appendix Table 2: In Appendix Fig. 2a-b and 5, we generate synthetic data for a 4-class problem with $p = 200$, $N_1 = 25$ (blue), $N_2 = 25$ (red), $N_1 = 25$ (green), $N_2 = 25$ (orange), $\delta = 0.5$, $\sigma = 0.075$, and cluster centroids 1,2 $\in \{-1.0, 1.0, -0.4, 0.4\}$.

In Appendix Figs. 2c-d, 2 and Table 2, we generate synthetic data for a 4-class problem with $p = 100$, $N_1 = 250$ (blue), $N_2 = 250$ (red), $N_3 = 250$ (green), $N_4 = 250$ (orange), $\delta = 0.5$, $\sigma = 0.075$, and cluster centroids 1,2,3,4,5,6,7,8,9,10 $\in \{-5.2, 5.2, -4.8, 4.8, -4.1, 4.1, -3.7, 3.7, -3.1, 3.1, -2.8, 2.8, -2.1, 2.1, -1.5, 1.5, -1.0, 1.0, -0.4, 0.4\}$.

Synthetic datasets 16–18. Appendix Figure 4: The data-generating parameters used for the datasets in Appendix Fig. 4a-b were 3 classes with $N_{\text{class}} = 50, \text{class} = 1, \ldots, 3$, $p = 20$, cluster centroids $\in \{-0.35, 0.0, 0.35\}$, and variable redundancy $\delta = 1$. In Fig. 4c, the data-generating parameters were 5 classes with $N_{\text{class}} = 20, \text{class} = 1, \ldots, 5$, $p = 50$, cluster centroids $\in \{-1.5, 1.5, 0.0, 0.01, -0.4\}$, and variable redundancy $\delta = 1$. For fitting PCMF, LL-PCMF, and P3CA, ρ was set to 1.0, γ was set to 2.0, the number of ADMM iterations per λ penalty was set to $K = 100$, and the number of nearest neighbors $N.N.$ was set to 25. For PCMF, we fit the model using problem ranks $1 \leq r \leq 10$ (Appendix Fig. 4a), and for LL-PCMF and P3CA, we fit the model using the $r = 1$ problem formulation (Appendix Fig. 4b). For the penalty path, we varied the value of λ along a path of 50 evenly spaced points in the interval $e^{[5,-20]}$ after initializing with $\lambda = \infty$ (10 iterations), such that the penalty decreased along the path of embedding solutions.

Synthetic dataset 19. Appendix Figure 3: We generate synthetic data for a 2-class problem with $p = 50$, $N_1 = 20$ (blue), $N_2 = 30$ (orange), $\delta = 1.0$, $\sigma = 0.75$, and cluster centroids 1,2 $\in \{-1, 1\}$ with two true slopes/principal directions per cluster.

Synthetic dataset 20. Appendix Figure 6: We generate synthetic data for a 4-class problem with $p = 20$, $N_1 = 100$ (blue), $N_2 = 50$ (orange), $N_3 = 20$ (red), $N_4 = 50$ (green), $\delta = 1.0$, $\sigma = 0.08$, and cluster centroids 1,2 $\in \{-0.35, -0.1, 0.05, 0.35\}$.

5.2 Real-World Datasets

NCI dataset. The **NCI Cancer Genomics dataset** consists of cDNA microarray gene expression levels in $p = 6,830$ genes measured in $N = 64$ cell lines from 13 cell types. We did not apply any preprocessing to the data, as the data had already been cleaned and prepared for use as a standard dataset. The data was originally prepared and released at http://genome-www.stanford.edu/nci60/ and is presently accessible on the (Hastie et al., 2009) book's website: https://hastie.su.domains/ElemStatLearn/datasets. The original data came from the (Ross et al., 2000) study. The cell classes and frequency per class are the following: 7 breast cancer cells, 5 central nervous system (CNS) cancer cells, 7 colon cancer cells, 1 K562B-repro (leukemia cell from a leukemia subtype cell line), 1 K562A-repro (leukemia cell from a leukemia subtype cell line), 6 leukemia cells, 1 MCF7A-repro (mamammary adenocarcinoma cell from a breast cancer subtype cell line), 1 MCF7D-repro (mamammary adenocarcinoma from a breast cancer subtype cell line), 8 melanoma cells, 9 non-small cell lung cancer (NSCLC) cells, 6 ovarian cancer cells, 2 prostate cancer cells, 9 renal cancer cells, and 1 unknown cancerous cell.

SRBCT dataset. The **SRBCT Cancer Genomics dataset** consists of cDNA microarray gene expression levels in $p = 2,318$ genes measured in $N = 88$ small, round blue-cell tumors (SRBCTs) of childhood samples from 4 cancer diagnostic categories. We did not apply any preprocessing to the data, as the data had already been cleaned and prepared for use as a standard dataset, however, we did combine the $N = 63$ training and $N = 25$ test set samples to maximize the sample size. The data was originally prepared and released at http://genome-www.stanford.edu/ and is presently accessible on the (Hastie et al., 2009) book's website: https://hastie.su.domains/ElemStatLearn/datasets. The original data came from the (Khan et al., 2001) study. The SRBCTs of childhood classes and frequency per class are the following: 11 Ewing family of tumors

(EWS: class 1), 29 rhabdomyosarcoma (RMS: class 2), 18 neuroblastoma (NB: class 3), 25 non- Hodgkin lymphoma (NHL/BL: class 4), and 5 unlabeled in the test set (class "NA").

Mouse dataset. The **Mouse Organ Cancer Genomics dataset** consists of single-cell RNA-sequencing in $p = 16,944$ genes measured in $N = 125$ mouse organ samples from 7 different mouse organs collected in the Tabula Muris study (Tabula Muris Consortium et al., 2018, Kopf et al., 2021). The organ classes and frequency per class are the following: 36 heart, 4 kidney, 37 large intestine, 8 liver, 8 lung, 20 spleen, 12 thymus. The 125 samples is a representative sample from the full dataset of $N = 6,232$ mouse organ samples. Data was scaled to be between 0 and 1 and genes were filtered to remove genes whose expression had low variance across rows (0.2 quantile) following previously published preprocessing steps for this dataset (Kopf et al., 2021).

Tumors dataset. The **Multiomics Cancerous Tumor dataset** consists of $p = 11,931$ multiomics measurements (concatenated measures from gene expression levels, DNA methylation, miRNA expression to obtain 1 data matrix) in tumor samples from $N = 142$ patients for 3 cancer diagnoses (glioblastoma multiforme (GBM), breast invasive carcinoma (BIC), and lung adenocarcinoma) from The Cancer Genome Atlas Program (TCGA) Research Network [https://www.cancer.gov/tcga] and curated by (Franco et al., 2021). The Tumors dataset classes and frequency per class are the following: 71 GBM cancer, 35 breast cancer, 36 lung cancer. The $N = 142$ samples is a representative sample from the full dataset of 424 patient samples. Gene expression levels, DNA methylation, miRNA expression measurements for each cancer type (GBM, BIC, and lung adenocarcinoma) were concatenated and then the three cancer types were merged such that all cancer types had the same features with no missing values. Data was scaled as following:

$$X_n = \frac{X_i - x_{min}}{x_{max} - x_{min}} \tag{53}$$

where X_i is the data for feature i while x_{max} and x_{min} are the minimum and maximum absolute value of the feature respectively. X_n is the normalized feature over rows. This followed previously published preprocessing steps for these datasets (Franco et al., 2021).

Tumors-Large dataset. The **Multiomics Cancerous Tumors-Large dataset** is the same dataset as the Tumors dataset, and is preprocessed in the same way, however with the full sample being used. It consists of consists of $p = 11,931$ single-view dataset (concatenated measures from gene expression levels, DNA methylation, miRNA expression to obtain 1 data matrix) in tumor samples from $N = 400$ patients in the training set and $N = 24$ patients in the test set for 3 cancer diagnoses (glioblastoma multiforme (GBM), breast invasive carcinoma (BIC), and lung adenocarcinoma). The Tumors dataset classes: GBM cancer, breast cancer, lung cancer, and the samples per class varied with the training/test set cross-validation fold.

Monkey-LGN dataset. The **Monkey-LGN Dataset** consists of $p = 45,768$ genes (expression) (X) measured from $N = 1,801$ cells of 2 cell types.

Mouse-LGN dataset. The **Mouse-LGN Dataset** consists of $p = 39,670$ genes (expression) (X) measured from $N = 1,818$ cells of 2 cell types.

MNIST dataset. The **MNIST dataset** was loaded from keras.datasets in Python, and we applied Y preprocessing step. We used six MNIST digit classes, 0,1,2,3,4,5, and the samples per class varied with the training/test set cross-validation fold.

MNIST Fashion dataset. The **MNIST fashion dataset** was downloaded from X, and we applied Y preprocessing step. We used six MNIST fashion classes, 0 (T-shirt/top), 1 (Trouser), 2 (Pullover), 3 (Dress), 4 (Coat), 5 (Sandal), and the samples per class varied with the training/test set cross-validation fold.

Human-ATAC dataset. The **Human-ATAC Dataset** consists of $p = 21,972$ chromatin profiles (X) from $30,480$ cells from 2 cell types.

COVID-19 dataset. The **COVID-19 multiomics dataset** was downloaded from (Shen et al., 2020). We used data samples corresponding to healthy subjects, moderate COVID-19 severity, and severe COVID-19 patients.

The dataset consisted of $p_X = 403$ metabolites and $p_X = 382$ proteins from $N = 45$ subjects ($N = 14$ healthy controls, $N = 18$ moderate COVID, $N = 13$ severe COVID) (Shen et al., 2020). We ran P3CA on using these X and Y datasets and separately CCA followed by K-means clustering for comparison. For comparison to other single-view clustering methods, we stacked the X and Y such that $XY = 45$ N x 785 p and used this XY as the input into each algorithm.

NCI (Multiview) dataset. The **NCI (Multiview) dataset** is the same dataset as the NCI dataset, however, we now split the dataset into X and Y such that: $X = 64$ N x $p_{[\text{variables } 1-1,000]}$ and $Y = 64$ N x $p_{[\text{variables } 1,001-1,100]}$ and ran P3CA on using these X and Y datasets and separately CCA followed by K-means clustering for comparison. For comparison to other single-view clustering methods, we stacked the X and Y such that $XY = 64$ N x $p_{[\text{variables } 1-1,100]}$ and used this XY as the input into each algorithm.

SRBCT (Multiview) dataset. The **SRBCT (Multiview) dataset** is the same dataset as the SRBCT dataset, however, we now split the dataset into X and Y such that: $X = 88$ N x $p_{[\text{variables } 1-1,000]}$ and $Y = 88$ N x $p_{[\text{variables } 1,001-1,100]}$ and ran P3CA on using these X and Y datasets and separately CCA followed by K-means clustering for comparison. For comparison to other single-view clustering methods, we stacked the X and Y such that $XY = 88$ N x $p_{[\text{variables } 1-1,100]}$ and used this XY as the input into each algorithm.

Mouse (Multiview) dataset. The **Mouse (Multiview) dataset** is the same dataset as the Mouse dataset, however, we now split the dataset into X and Y such that: $X = 125$ N x $p_{[\text{variables } 1-1,000]}$ and $Y = 125$ N x $p_{[\text{variables } 1,001-1,100]}$ and ran P3CA on using these X and Y datasets and separately CCA followed by K-means clustering for comparison. For comparison to other single-view clustering methods, we stacked the X and Y such that $XY = 125$ N x $p_{[\text{variables } 1-1,100]}$ and used this XY as the input into each algorithm.

Tumors (Multiview) dataset. The **Tumors (Multiview) dataset** is the same dataset as the Tumors dataset, however, we now split the dataset into X and Y such that: $X = 142$ N x $p_{[\text{variables } 1-1,000]}$ and $Y = 142$ N x $p_{[\text{variables } 1,001-1,100]}$ and ran P3CA on using these X and Y datasets and separately CCA + K-means clustering for comparison. For comparison to other single-view clustering methods, we stacked the X and Y such that $XY = 142$ N x $p_{[\text{variables } 1-1,100]}$ and used this XY as the input into each algorithm.

Autism spectrum disorder (ASD) dataset. The **autism spectrum disorder (ASD) dataset** consists of $p_{_X} = 3$ clinical symptoms and $p_{_Y} = 20$ resting state functional connectivity (RSFC) features measured from resting state functional MRI (rsfMRI) neuroimaging in $N = 299$ patients with ASD (top 20 RSFC correlated with three clinical symptoms; see Appendix and (Buch et al., 2023, Drysdale et al., 2017, Grosenick et al., 2019) for feature selection methods). The three clinical symptoms are verbal IQ (VIQ), ADOS-2 social affect CSS, and ADOS-2 repetitive behaviors and restricted interests (RRB) CSS. Datasets were collected from subjects as part of the ABIDE studies (Martino, 2014, 2017). The ADOS-2 is the Autism Diagnostic Observation Schedule-Second Edition CSS, a standardized observational scale for diagnosing ASD. CSS stands for the calibrated severity score.

Following standard preprocessing of the rsfMRI (Buch et al., 2023, Drysdale et al., 2017, Grosenick et al., 2019, Satterthwaite et al., 2012), we calculated the RSFC matrices for each subject by the Pearson correlation between 247 regions of interest (ROIs) from the Power brain atlas (Power et al., 2011). We next performed feature selection based on previous methods (Buch et al., 2023, Drysdale et al., 2017, Grosenick et al., 2019) by calculating the Spearman correlation between each RSFC feature ($p = 30,381$ unique RSFC) and each clinical symptom in 1,000 subsamples of 95% of the subjects ($N = 284$) and ranked RSFC features by the number of subsamples in which the RSFC had a significant correlation ($p < 0.05$) to one of the clinical symptoms. This rank list represented the relative importance of each RSFC feature to predicting clinical symptoms in ASD. We selected the top 20 RSFC from this rank list. For each view (X and Y), we add a column of ones as a free variable in which the U and V coefficients can capture differences in cluster means. Thus the input into the analysis included the $p_X = 4$ (three clinical symptoms concatenated with a column of $N = 299$ ones as an intercept term) and $p_Y = 21$ (20 most predictive RSFC features concatenated with a column of $N = 299$ ones as an intercept term) in $N = 299$ patients with ASD.

All human neuroimaging and behavioral data from the ABIDE I and II datasets is anonymized, there is no protected health information included, and the datasets are publicly available with approval. Protocols for human subject research in the ABIDE datasets are included in the study details for each of the 17 study sites for ABIDE

I and 19 study sites for ABIDE II (see http://fcon_1000.projects.nitrc.org/indi/abide). Thus we did not collect any data using human subjects for this study, and all information about IRB approval and participant compensation can be found in the original datasets collected by the ABIDE I and II consortia. As we using rsfMRI data and the behavioral measures included are from standardized behavioral and clinical scales (intelligence quotient and ADOS-2), participants were not shown text instructions during MRI scanning.

Palmer Penguin (Multiview) dataset. The **Palmer Penguin Dataset** consists of four body measurements in $N = 342$ penguins from three species. For illustration of P3CA, we split the four measurements into two views: X with bill length and bill depth; Y with flipper length and body mass. The Palmer Penguin dataset classes and frequency per class are the following: 151 Adélie penguins, 68 Chinstrap penguins, 123 Gentoo penguins.

5.3 Protein-Protein Interaction Network

To assess interpretation of the cluster-specific P3CA embeddings for the COVID-19 dataset, we performed a graph-based network analyis called Protein Protein Interaction (PPI) network analysis using the top 25 proteins whose expression/abundance was most correlated (Spearman correlation) with the P3CA embedding for each severity cluster. We converted protein IDs to standardized HGNC gene names and separately input each 25 gene name list into the online NetworkAnalyst (https://www.networkanalyst.ca/) platform (1 per PPI model/cluster) (Xia et al., 2014, 2015, Zhou et al., 2019). The seeds and interaction partners were used to build a zero-order PPI subnetwork, meaning each PPI only included proteins that were from the input 25 protein/gene seeds. We calculated the degree (number of connections) for each protein/gene in the PPI networks and used Enrichr (https://maayanlab.cloud/Enrichr/) (Chen et al., 2013, Kuleshov et al., 2016, Xie et al., 2021) to identify the overlap between proteins in the PPI networks known to be associated with COVID-19.

5.4 Dendrograms and Interpretation

We construct PCMF dendrograms following the procedure outlined in Main Text §3.2. For comparison to the result when using a two-step embedding then clustering method, we compare the results in Main Text Figs. 2 and 3 to an embedding method (PCA or CCA) followed sequentially by agglomerative hierarchical clustering. We cut the hierarchical clustering dendrogram at the true number of clusters for comparison and then calculate and report the clustering accuracy from the two-step procedure.

5.5 Hyperparameters

PCMF, LL-PCMF, and P3CA: How is the rank tuned?

We have not tuned the rank as we focused on the rank 1 case for all experiments where we compare the rank 1 globally linear PCMF against the rank 1 locally linear LL-PCMF formulation. In the consensus version of PCMF, we set the rank to be of higher order to increase flexibility of the clustering solution and to accommodate "spurious" local nonlinearities that may be present in the huge-sample datasets. Future work can explore the tuning of the rank, by incorporating rank as a hyperparameter to a standard cross-validation scheme with train/validation/test set folds.

Cluster comparison methods: What hyperparameters did we use?

In comparison, for the deep clustering methods, the Louvain, and the Leidein method, we implemented a finer hyperparameter grid for tuning as detailed below. With additional hyperparameter tuning on a finer grid, we expect our PCMF and P3CA model may be further optimized.

For **Deep Embedding Clustering**, we tuned the following hyperparameters: batch size (15, 30), finetune iterations (100, 1000), iterations for layerwise pretraining (100, 1000), and maximum iterations for clustering (100, 200). Layer sizes and other parameters were set to the defaults in the model code from https://github.com/fferroni/DEC-Keras.

For **Improved Deep Embedding Clustering**, we tuned the following hyperparameters: batch size (15, 30), pretraining epochs (100, 1000) and training epochs (100, 1000). Layer sizes and other parameters were set to the defaults in the model code from https://github.com/dawnranger/IDEC-pytorch.

For **CarDEC**, we tuned the following hyperparameters: number of neighbors (5, 10, 15, 20, 25) and number of top genes (100, 500).

For **Louvain** and **Leidein**, we tuned the number of neighbors (5, 10, 15, 20, 25).

The optimal hyperparameters for these deep clustering and Louvain and Leidein methods were as follows (*N.N.* indicates number of nearest neighbors):

1. For the **NCI dataset**: Leiden with 5 *N.N.*; Louvain with 15 *N.N.*; DEC with batch size 15, finetune iterations 100, layerwise iterations 100, cluster iterations 200; IDEC with batch size 30, pretrain iterations 100, and train iterations 1000; CarDEC with 5 *N.N.* and 100 top genes.
2. For the **SRBCT dataset**: Leiden tied with 10, 20, or 25 *N.N.*; Louvain with 15 *N.N.*; DEC with batch size 15, finetune iterations 1000, layerwise iterations 1000, cluster iterations 100; IDEC with batch size 30, pretrain iterations 100, and train iterations 100; CarDEC with 10 *N.N.* and tied with 100 or 200 top genes.
3. For the **Mouse Organs dataset**: Leiden tied with 20 or 25 *N.N.*; Louvain with 15 *N.N.*; DEC with batch size 15, finetune iterations 100, layerwise iterations 100, cluster iterations 100; IDEC with batch size 30, pretrain iterations 100, and train iterations 100; CarDEC with 15 *N.N.* and 100 top genes.
4. For the **Tumors dataset**: Leiden tied with 15, 20, or 25 *N.N.*; Louvain tied with 20 or 25 N.N.; DEC tied with batch size 15 or 30, finetune iterations 100 or 1000, layerwise iterations 100 or 1000, cluster iterations 100 or 200; IDEC with batch size 15, pretrain iterations 100, and train iterations 100; CarDEC with 20 *N.N.* and 200 top genes.
5. For the **COVID-19 dataset**: Leiden tied with 20 or 25 *N.N.*; Louvain with 10 *N.N.*; DEC with batch size 15, finetune iterations 1000, layerwise iterations 100, cluster iterations 200; IDEC with batch size 30, pretrain iterations 100, and train iterations 100; CarDEC with 5 *N.N.* and 200 top genes.
6. For the **Tumors-large dataset**: Leiden with 40 *N.N.*; Louvain with 40 *N.N.* IDEC with batch size 15, pretrain iterations 100, and train iterations 100.
7. For the **MNIST dataset**: Leiden with 10 *N.N.*; Louvain with 40 *N.N.*; IDEC with batch size 15, pretrain iterations 100, and train iterations 100.
8. For the **FashionMNIST dataset**: Leiden with 40 *N.N.*; Louvain with 40 *N.N.*; IDEC with batch size 15, pretrain iterations 100, and train iterations 100.
9. For the **Synthetic dataset**: Leiden with 40 *N.N.*; Louvain with 25 *N.N.*

References

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

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011b.

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.

E. Y. Chen, C. M. Tan, Y. Kou, Q. Duan, Z. Wang, G. V. Meirelles, N. R. Clark, and A. Ma'ayan. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1):128, Apr. 2013.

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

R. Couillet and Z. Liao. *Random matrix methods for machine learning*. Cambridge University Press, Cambridge, England, Aug. 2022.

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.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Am. Stat. Assoc.*, 97(457):77–87, Mar. 2002.

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

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.

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.

J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *Ann. Appl. Stat*, 1(2): 302–332, Dec. 2007.

D. Gabay and B. Mercier. A dual algorithm for the solution of nonlinear variational problems via finite element approximation. *Comput. Math. Appl.*, 2(1):17–40, Jan. 1976.

R. Glowinski and A. Marroco. Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *[Revue fr. autom. inform. rech. opér., Anal. numér.*, 9(R2):41–76, 1975.

A. Goncalves, X. Liu, and A. Banerjee. Two-block vs. multi-block ADMM: An empirical evaluation of convergence. *arXiv*, July 2019.

L. Grosenick, B. Klingenberg, K. Katovich, B. Knutson, and J. E. Taylor. Interpretable whole-brain prediction analysis with GraphNet. *Neuroimage*, 72:304–321, May 2013.

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.

M. G. G'Sell, J. Taylor, and R. Tibshirani. Adaptive testing for the graphical lasso. *arXiv*, July 2013.

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.

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.

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

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.

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.

U. Kruger and S. Joe Qin. Canonical correlation partial least squares. *IFAC Proceedings Volumes*, 36(16): 1603–1608, Sept. 2003.

M. V. Kuleshov, M. R. Jones, A. D. Rouillard, N. F. Fernandez, Q. Duan, Z. Wang, S. Koplev, S. L. Jenkins, K. M. Jagodnik, A. Lachmann, M. G. McDermott, C. D. Monteiro, G. W. Gundersen, and A. Ma'ayan. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.*, 44(W1):W90–7, July 2016.

R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Ann. Stat.*, 42(2): 413–468, Apr. 2014.

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

D. A. Martino. The Autism Brain Imaging Data Exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry*, 19(6):659, 2014.

D. A. Martino. Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci Data*, 4:170010, 2017.

E. Min, X. Guo, Q. Liu, G. Zhang, J. Cui, and J. Long. A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514, 2018.

N. Parikh and S. Boyd. Proximal algorithms. *Found. Trends Optim.*, 1(3):127–239, Jan. 2014.

J. D. Power, A. L. Cohen, S. M. Nelson, G. S. Wig, K. A. Barnes, J. A. Church, A. C. Vogel, T. O. Laumann, F. M. Miezin, B. L. Schlaggar, and S. E. Petersen. Functional network organization of the human brain. *Neuron*, 72(4):665–678, Nov. 2011.

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.

T. D. Satterthwaite, D. H. Wolf, J. Loughead, K. Ruparel, M. A. Elliott, H. Hakonarson, R. C. Gur, and R. E. Gur. Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage*, 60(1):623–632, Mar. 2012.

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.

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.

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.

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

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stat. Sci.*, 18(1):104–117, 2003.

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.

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.

J. Xia, M. J. Benner, and R. E. W. Hancock. NetworkAnalyst–integrative approaches for protein-protein interaction network analysis and visual exploration. *Nucleic Acids Res.*, 42(Web Server issue):W167–74, July 2014.

J. Xia, E. E. Gill, and R. E. W. Hancock. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. *Nat. Protoc.*, 10(6):823–844, June 2015.

Z. Xie, A. Bailey, M. V. Kuleshov, D. J. B. Clarke, J. E. Evangelista, S. L. Jenkins, A. Lachmann, M. L. Wojciechowicz, E. Kropiwnicki, K. M. Jagodnik, M. Jeon, and A. Ma'ayan. Gene set knowledge discovery with enrichr. *Curr Protoc*, 1(3):e90, Mar. 2021.

G. Zhou, O. Soufan, J. Ewald, R. E. W. Hancock, N. Basu, and J. Xia. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.*, 47(W1):W234–W241, July 2019.