Supervised convex clustering

Minjie Wang¹ | Tianyi Yao² | Genevera I. Allen³

¹School of Statistics, University of Minnesota, Minneapolis, Minnesota, USA
²Department of Statistics, Rice University, Houston, Texas, USA
³Department of Electrical and Computer Engineering, Statistics, and Computer Science, Rice University and Jan and Dan Duncan Neurological Research Institute, Baylor College of Medicine, Houston, Texas, USA

Abstract
Clustering has long been a popular unsupervised learning approach to identify groups of similar objects and discover patterns from unlabeled data in many applications. Yet, coming up with meaningful interpretations of the estimated clusters has often been challenging precisely due to their unsupervised nature. Meanwhile, in many real-world scenarios, there are some noisy supervising auxiliary variables, for instance, subjective diagnostic opinions, that are related to the observed heterogeneity of the unlabeled data. By leveraging information from both supervising auxiliary variables and unlabeled data, we seek to uncover more scientifically interpretable group structures that may be hidden by completely unsupervised analyses. In this work, we propose and develop a new statistical pattern discovery method named supervised convex clustering (SCC) that borrows strength from both information sources and guides towards finding more interpretable patterns via a joint convex fusion penalty. We develop several extensions of SCC to integrate different types of supervising auxiliary variables, to adjust for additional covariates, and to find biclusters. We demonstrate the practical advantages of SCC through simulations and a case study on Alzheimer’s disease genomics. Specifically, we discover new candidate genes as well as new subtypes of Alzheimer’s disease that can potentially lead to better understanding of the underlying genetic mechanisms responsible for the observed heterogeneity of cognitive decline in older adults.

KEYWORDS
convex clustering, exponential family, generalized linear model deviance, interpretable clustering, supervised clustering

1 | INTRODUCTION
Clustering is an unsupervised learning approach that seeks to find groups of objects which are similar to each other. Despite successes in applying clustering in many fields such as genomics, online advertising, and text mining, coming up with meaningful interpretations of the estimated clusters has often been challenging precisely due to its unsupervised nature. Currently, in practice, most people cluster data in a fully unsupervised manner and then interpret the clustering results via some outcomes of interest or other meta-data that help to validate the clusters. We call these “supervising auxiliary variables”. Our goal is to use the supervising auxiliary variables as part of the clustering procedure itself to help guide towards finding more accurate and interpretable clusters. Let us consider...
our motivating case study on the clinical genomics of Alzheimer’s disease (AD), which will be discussed in more detail in Section 4. All individuals, in this case study, experience cognitive decline as they age, yet cognitive abilities of some subjects decline at a much faster rate than others, and there is a large degree of heterogeneity in cognitive skills. Understanding such heterogeneity in cognitive decline can better elucidate the underlying genetic mechanisms responsible for AD and other dementias. However, apart from a handful of well-known genes such as APOE, little is known about the genomics of AD and of cognitive decline in older adults. It is common for people to study this by clustering the subjects and validating the results using additional information such as the clinical diagnosis or cognitive test scores. However, compelling evidence of genetic subtypes for AD and cognitive decline has not been found. We propose to use the additional meta information such as the cognitive test scores or clinical diagnosis directly to help us find better and more interpretable clusters, and hence, shed new light on the genetic basis responsible for the onset of AD and dementia. In many other genomics studies, additional clinical information or survival times are available and can thus be used as supervising auxiliary variables to guide clustering.

Yet, making use of the supervising auxiliary variables to help guide towards finding groups presents several major challenges. First, due to human subjectivity and measurement errors, these clinical outcomes are noisy and thus cannot be fully trusted as ground-truth outcomes or labels. Specifically, we do not know how much supervision we should get from these noisy supervising auxiliary variables when clustering unlabeled data. Second, these clinical outcomes can be of different data types. For example, diagnostic opinions assigned by biologists can be categorical while survival time is censored data.

To address these challenges, in this paper, we seek to leverage information from both supervising auxiliary variables, usually of different types, and unlabeled data to uncover more scientifically interpretable group structures that may be hidden in completely unsupervised analyses of data. Our approach is distinct from supervised learning, which treats these outcomes as ground truth to make scientific discoveries, as the supervised approaches fail to exploit the unlabeled data to uncover group structures. In addition, these supervising auxiliary variables are different from the outcomes or labels in supervised learning in that they are largely noisy and thus cannot be fully trusted. Our method is also distinct from unsupervised approaches as we make better use of these potentially meaningful supervising auxiliary variables to understand the true underlying group structures. Although these supervising auxiliary variables are not true labels or outcomes themselves, they have a loose relationship with the group structure in the data and hence indicate some forms of observed heterogeneity of the unlabeled data.

Though supervised clustering has not been widely studied, there is a plethora of literature on semi-supervised clustering specific to the nature of the outcome variables. For partially labeled data, Basu et al. (2002) proposed to modify the objective function of $k$-means to compute initial cluster centroids by incorporating such labels. In this case, labeled observations are always assigned to their known cluster. However, such approaches usually assume the labels are perfect and require prior knowledge of the total number of clusters. In other scenarios, people incorporate prior information on pairwise (must-link or cannot-link) constraints that dictate whether two data points must be clustered in the same group or not. To take those constraints into account, Basu et al. (2004) and Xing et al. (2002) modified the objective function of existing clustering methods or the distance metric in the distance-based clustering method. Still, those methods require that the presumed knowledge of the constraints is correct. On the other hand, many have tried to improve the interpretability of clustering results by only using features related to some supervised outcome. Specifically, Bair and Tibshirani (2004) and Koestler et al. (2010) discarded or down-weighted unimportant features by univariate filtering associated with the outcome and then performed clustering on the meaningful features. However, such methods might neglect features that are weakly associated with the outcome variable but differ across clusters. Above all, these semi-supervised clustering approaches make full use of the noisy outcome variable without adjustment. Yet, our goal is to leverage information from both supervising auxiliary variables, which may be imperfect, and unlabeled data to obtain more interpretable group structures.

Another line of work focuses on semi-supervised classification which uses both labeled and unlabeled data to improve the performance of classifiers. We provide related literature on semi-supervised classification in Web Appendix G. Our method is fundamentally different from both semi-supervised classification or classification with noisy labels methods in that the major task of these methods is supervised learning (prediction), while our goal is to find groups using the supervising auxiliary variables.

We seek to develop a unified, convex formulation of supervised clustering based on increasingly popular convex clustering methods. Pelckmans et al. (2005), Lindsten et al. (2011), and Hocking et al. (2011) studied a convex formulation of clustering that achieves agglomeration through a convex fusion penalty. Due to this convex formulation, it enjoys nice statistical properties such as global optimal solutions, stable solutions to small perturbations of data (Chi et al., 2017; Pelckmans et al., 2005) and statistical consistency (Radchenko & Mukherjee, 2017;
Tan & Witten, 2015). Recently, to address the expensive computation of convex clustering, Chi and Lange (2015) and Weylandt et al. (2020) developed fast and efficient algorithms to solve the convex clustering problem and yield full regularization paths. Further, convex clustering has been extended to many applications such as convex biclustering (Chi et al., 2017), which allows for clustering features simultaneously, and closely related to our work, recently Wang and Allen (2021) adopted the convex clustering approach to perform integrative clustering for high-dimensional mixed, multiview data. The objective function of convex clustering bears a resemblance to the general model and formulation [\text{g}](\cdot) penalty on the differences of the neighbor-differences of the neighboring coefficients. Convex clustering with \ell_1-norm fusion is a special case of the fused lasso.

In this paper, we propose and develop a new statistical pattern discovery method named supervised convex clustering (SCC) that borrows strength from both the unlabeled data and supervising auxiliary variables to find more interpretable patterns. Specifically, we develop an optimization problem defining our method that consists of three parts: an unsupervised loss for the unlabeled data, a supervised loss that incorporates the supervising auxiliary variable, and a joint convex fusion penalty that forces the group structure of the unlabeled data and the supervising auxiliary variable to be the same. Our method, to the best of our knowledge, is the first to perform supervised clustering that directly uses supervising auxiliary variables to help cluster unlabeled data. The novelty of the paper is how we translate this novel idea of leveraging supervising auxiliary variables for achieving more interpretable clusters into a coherent optimization problem that is convex and solvable.

2 | SUPERVISED CONVEX CLUSTERING METHOD

In this section, we propose and develop our SCC method for different types of supervising auxiliary variables. Then we discuss some practical considerations for applying our method and develop an adaptive approach to adjust for additional covariates.

2.1 | General model and formulation

Let \((y_i, X_i)\) denote the pair of supervising auxiliary variable \(y_i\) and feature vector \(X_i \in \mathbb{R}^p\) for the \(i\)th observation, \(i \in \{1, \ldots, n\}\). Let \((\theta_i, U_i)\) denote the corresponding pair of supervising auxiliary variable centroid \(\theta_i\) and data centroid \(U_i \in \mathbb{R}^p\) for the \(i\)th observation. Let \(Z_i \in \mathbb{R}^d\) denote the additional covariates associated with the supervising auxiliary variable for the \(i\)th observation and \(\beta \in \mathbb{R}^d\) is the corresponding vector of coefficients. Define \(g(\cdot)\) to be the appropriate link function for a generalized linear model (GLM) whose exact form depends on the data type of the supervising auxiliary variable \(y\) (continuous, skewed-continuous, binary, and count-valued, among others). For example, if \(y\) is count-valued data, \(g(\cdot)\) can be the log-link. Define \(C : i \rightarrow k\) to be a function that maps from the observation indices \(i \in \{1, \ldots, n\}\) to cluster labels \(k \in \{1, \ldots, K\}\). Then for \(i \in \{1, \ldots, n\}\), we consider the following data-generating model:

\[
X_i = U_i + E_i, \quad E_i \sim \text{MVN}(0_p, \sigma^2 I), \quad g(E[y_i | Z_i]) = \theta_i + Z_i^T \beta, \quad \sigma^2, \quad \beta \in \mathbb{R}^d,
\]

This model assumes the unlabeled data follows a group mean (centroid) plus noise model. The supervising auxiliary variable, adjusted for covariates \(Z_i\), follows a GLM whose mean has the same group structure as the unlabeled data.

We propose to fit this model by formulating a convex optimization problem based on convex clustering. Let \(\ell(\cdot)\) be the negative log-likelihood or loss function for the particular GLM associated with \(g(\cdot)\). Our SCC method is hence the solution to the following optimization problem:

\[
\min_{U \in \mathbb{R}^{n \times p}, \theta \in \mathbb{R}^K, \beta \in \mathbb{R}^d} \\pi_x \cdot \frac{1}{2} \sum_{i=1}^n ||X_i - U_i||_2^2 + \pi_y \cdot \sum_{i=1}^n \ell(y_i; \theta_i + Z_i^T \beta) + \lambda \sum_{1 \leq i < j \leq n} w_{ij} \left[ ||\theta_i - \theta_j||_2^2 + ||U_i - U_j||_2^2 \right], \quad (1)
\]

Here, \(\pi_x\) and \(\pi_y\) are fixed inputs by the user in advance; \(\lambda\) is a nonnegative tuning parameter; and, the weights \(w_{ij}\) are nonnegative user-specific fixed inputs.

Our optimization problem can be thought of as an extension of convex clustering that incorporates supervised data. One way to interpret this is that we have a loss function for the unlabeled data, a loss function for the supervising auxiliary variable, and a new joint convex fusion penalty that connects the supervised and unsupervised parts. Specifically, we employ a joint group-lasso fusion penalty on the concatenated centroid \([\theta \quad U]\) that forces the group struc-
ture (cluster assignment) of the ith row of U to be the same as that of Θ. Similar to convex clustering, our joint group-lasso-type fusion penalty encourages the differences in the rows to be shrunk towards zero, inducing a clustering behavior. Here, λ is a positive tuning parameter that regulates both the cluster assignment and number of clusters. When λ equals zero, each observation forms its own cluster centroid. As λ increases, the fusion penalty encourages the rows of concatenated centroid [Θ U] to merge together, forming clusters. We say that subjects with the same centroids belong to the same cluster, which means, Xl and Xj have the same cluster membership if Uj = Ul and Θj = Θl.

As λ is sufficiently large, all the rows of concatenated centroid [Θ U] coalesce to a single cluster centroid. Our joint fusion penalty is novel as it puts together the centroids for both the unlabeled data and the supervising auxiliary variable, forcing them to have the same group structure. In this way, our method borrows strength from both information sources and yields the same cluster assignment for similar observations. The weight wij, which manifests pairwise affinity, is a user-specific input which will be discussed in detail in Section 2.5.

The two loss functions, \( \frac{1}{2} \sum_{i=1}^{n} ||X_i - U_i||_2^2 \) and \( \ell(y_i; \theta_i + Z_i^T \beta) \), are usually of different scales. If the value of one loss function dominates the other, the clustering assignment will favor the data source with a larger loss function value. Therefore, we weight the two losses by \( \pi_X \) and \( \pi_Y \), which are deterministic parameters for fixed data, to ensure that the losses of two data sources are evaluated at the same scale in the objective function. Inspired by the likelihood-ratio test statistic, we propose setting \( \pi_X \) and \( \pi_Y \) to be inversely proportional to the null deviance evaluated at the loss-specific center, that is, \( \pi_X = \frac{1}{\frac{1}{2} ||X_i - X_i||_2^2} \), \( \pi_Y = \frac{1}{\ell(y_i, y_i)} \), so that the losses are measured at the same scale. Here \( \hat{y} \) denotes the loss-specific center for loss \( \ell(\cdot) \) as discussed in Wang and Allen (2021).

Note that the null deviance \( \ell(y, \hat{y}) \) equals the deviance under the null hypothesis that there is only one cluster. Suppose we remove the second term in the objective function (\( \pi_Y = 0 \)), we get a fully unsupervised method. Similarly, if we remove the first term (\( \pi_X = 0 \)), we perform convex clustering on the supervising auxiliary variable alone. We further discuss the choice of parameters in Web Appendix F.

We employ different loss functions to account for the unlabeled data and supervise auxiliary variables of different data types. Notice that here we assume the data matrix X to follow a Gaussian distribution and use Euclidean distances as the loss function, but one could easily generalize it to any convex losses for non-Gaussian data X.

The general loss \( \ell(\cdot) \) is a convex function whose specific form depends on the data type of y. For example, \( \ell \) can be the negative log-likelihood for any common GLMs such as Gaussian, logistic, log-linear (Poisson), and negative binomial. For a supervising auxiliary variable which is categorical or survival data, the general form above does not apply and we need some minor changes to the formulation. We specify the categorical y below as a special case and discuss the case when y is censored survival time in Web Appendix D.

2.2 | Special case: Categorical supervising auxiliary variable

We model a categorical supervising auxiliary variable with K classes using the multinomial loss. To facilitate this, we first transform the supervising auxiliary variable into dummy variables \( Y \in \mathbb{R}^{n \times K} \), where \( Y_{ik} = 1 \) if subject i belongs to the kth class.

By construction, we employ the negative log-likelihood of multinomial distribution as loss \( \ell(\cdot) \) and similarly denote \( \Theta \in \mathbb{R}^{n \times K} \) as the centroid matrix for a supervising auxiliary variable \( Y \); this gives SCC for categorical supervising auxiliary variables:

\[
\min_{U \in \mathbb{R}^{n \times K}, \Theta \in \mathbb{R}^{n \times K}, \beta \in \mathbb{R}^d} \pi_X \cdot \frac{1}{2} \sum_{i=1}^{n} ||X_i - U_i||_2^2 + \pi_Y \cdot \sum_{i=1}^{n} \left\{ \sum_{k=1}^{K} -Y_{ik}(\theta_{ik} + z_i^T \beta_k) + \log(\sum_{k=1}^{K} e^{\theta_{ik} + z_i^T \beta_k}) \right\} \\
+ \lambda \sum_{1 \leq i < j \leq n} w_{ij} ||[\Theta_i - U_i] - [\Theta_j - U_j]||_2^2.
\]

We enforce a joint fusion penalty on the rows of concatenated centroid [Θ U] to yield a shared group structure between two sources. Note as in the regular multinomial regression problem, this parameterization of \( \beta_k \) is not identifiable as the value of the objective function would not change if we replace \( \beta_k \) with \( \beta_k + c \) for all \( k \). To address this issue, we add the constraint \( \sum_{k=1}^{K} \beta_k = 0 \) as discussed in J. Zhu and Hastie (2004).

2.3 | Supervised convex biclustering

To allow for grouping observations and features simultaneously, we extend our method to supervised convex biclustering based on the approach discussed by Chi et al. (2017). The supervised convex biclustering problem can be
formulated as follows:

$$\minimize_{\mathbf{U} \in \mathbb{R}^{n \times p}, \boldsymbol{\theta} \in \mathbb{R}^n, \mathbf{Z} \in \mathbb{R}^{p \times d}} \pi_X \cdot \frac{1}{2} \sum_{i=1}^{n} \| \mathbf{X}_i - \mathbf{U}_i \|^2_F + \pi_Y \cdot \sum_{j=1}^{n} \ell(y_j; \mathbf{z}_j) + Z \cdot \mathbf{z} + \lambda \sum_{1 \leq i \leq l \leq n} w_{ij} \| \mathbf{U}_i \|_2 + \lambda \sum_{1 \leq j \leq p} \tilde{w}_{jj} \| \mathbf{U} \|_2$$

subject to \( \mathbf{D} \cdot \mathbf{U} - \mathbf{V} = \mathbf{0} \).

To facilitate the constraints above in matrix form, we concatenate the supervising auxiliary variable centroid vector \( \mathbf{D} \in \mathbb{R}^{n \times p} \) and the data centroid matrix \( \mathbf{U} \in \mathbb{R}^{n \times (p+1)} \) into the aggregated centroid matrix \( \mathbf{D} \mathbf{U} \in \mathbb{R}^{n \times (p+1)} \). We then introduce an auxiliary variable \( \mathbf{V} = [\mathbf{V}_\theta \; \mathbf{V}_U] \in \mathbb{R}^{(n \times (p+1))} \) containing the pairwise differences between connected rows of the aggregated centroid matrix \( \mathbf{D} \mathbf{U} \). The constraints can now be written as \( \mathbf{D} \mathbf{U} - \mathbf{V} = \mathbf{0} \).

Also, we replace the squared Euclidean distances with the squared Frobenius norm to facilitate the matrix-version algorithm. We thus obtain the augmented Lagrangian and apply the multi-block ADMM to solve our SCC problem.

Further, note that both the \( \mathbf{D} \) and \( \mathbf{Z} \) subproblems generally do not have analytical closed-form solutions for an arbitrary loss function \( \ell \). Hence we need to apply an inner optimization routine with nested iterative updates to solve the subproblem until full convergence, which is computationally intensive. To address this and speed up computation, we adopt the generalized multi-block ADMM with an inexact subproblem approach by Wang and Allen (2021) and take a one-step descent update to solve the subproblem approximately. For a differentiable loss \( \ell \), we take a one-step gradient descent update by applying a linearized multi-block ADMM to the \( \mathbf{D} \) or \( \mathbf{Z} \) subproblem for each iteration. This gives Algorithm 1, a multi-block ADMM algorithm to solve supervised convex clustering with a differentiable loss. Similarly, for a non-differentiable distance-based loss \( \ell \), we introduce a new block for the nonsmooth function \( \ell \) and apply a multi-block ADMM with simple closed-form solutions for each primal variable update. In Algorithm 1, the constant \( \rho \) is the augmented Lagrangian parameter in ADMM; we find setting \( \rho \) to be 1 works well in practice. The step size \( t_k \) is chosen via backtracking. The proximal mapping of function \( h \) is defined as \( \text{prox}_{h; \cdot}(\mathbf{x}) = \arg \min_{\mathbf{z}} \frac{1}{2} \| \mathbf{x} - \mathbf{z} \|_2^2 + h(\mathbf{z}) \).

**Proposition 1.** If \( \ell \) is convex, Algorithm 1 converges to a global solution. In addition, if \( \ell \) is strictly convex, it converges to the unique global solution.

Proposition 1 is an extension of Theorem 4 of Wang and Allen (2021) and guarantees the convergence of multi-block ADMM using inexact subproblem approximations. Similarly, to solve the supervised convex biclustering...
Algorithm 1 Multi-block ADMM algorithm for supervised convex clustering with a differentiable loss $\ell$

\[
\text{while not converged do}
\]

\begin{align*}
(i) \quad & U^{(k+1)} = (\pi X \cdot I + \rho D^T D)^{-1} (\pi X + \rho D^T (V^{(k)}_U - Q^{(k)}_U)) ; \\
(ii) \quad & \theta^{(k+1)} = \theta^{(k)} - t_k (\pi y \cdot \nabla \ell(\gamma; \theta^{(k)} + Z \beta^{(k)}) + \rho D^T (D \theta^{(k)} - V^{(k)}_\theta + Q^{(k)}_\theta)) ; \\
(iii) \quad & \beta^{(k+1)} = \beta^{(k)} - t_k \nabla \ell(\gamma; \theta^{(k+1)} + Z \beta^{(k)}) ; \\
(iv) \quad & V^{(k+1)} = \text{prox}_{\lambda/\rho P(\cdot;\omega)} (D \left[ \theta^{(k+1)} \cdot U^{(k+1)} + Q^{(k)} \right]) ; \\
(v) \quad & Q^{(k+1)} = Q^{(k)} + (D \left[ \theta^{(k+1)} \cdot U^{(k+1)} \right] - V^{(k+1)}) ;
\end{align*}

\]

problem, we apply a multi-block ADMM and provide Algorithm 1 in Web Appendix A.

2.5 Practical issues

In this section, we address some practical issues of applying our methods to real data. First, we show how to choose the regularization parameter $\lambda$. Then, we discuss the choice of weights and tuning parameters for the level of supervision. Moreover, we introduce an adaptive method to adjust for additional covariates.

2.5.1 Choice of tuning parameter $\lambda$

As mentioned, the tuning parameter $\lambda$ regulates both the number of clusters and the cluster assignments. The same type of procedures people use to choose the regularization parameter $\lambda$ for other convex clustering methods work here. For example, J. Wang (2010) and Fang and Wang (2012) proposed stability selection based approaches; Chi et al. (2017) proposed hold-out validation; Tan and Witten (2015) considered information criterion based approaches.

2.5.2 Choice of weights and level of supervision

In practice, the choice of fusion weights has been shown to play an important role in computational efficiency and clustering quality. Chi and Lange (2015), Hocking et al. (2011), and Chi et al. (2017) have shown that setting weights inversely proportional to the distances between two observations yields superior performance. On the other hand, enforcing sparse weights reduces computational cost and improves clustering quality. Given these two, the most commonly used weights choice for convex clustering is the $k$-nearest-neighbors method with a Gaussian kernel.

The challenge here is that the unlabeled data $X$ and supervising auxiliary variable $y$ are measured from different sources and can be of different types; thus the Gaussian kernel with Euclidean distances is not an appropriate distance metric in this case. To measure the dissimilarity of two subjects measured in different data types, we adopt the Gower distance (Gower, 1971), which is a commonly used distance metric for mixed types of data and has been shown to obtain superior performance compared with other distance metrics (Ali & Massmoudi, 2013; Hummel et al., 2017; Wang & Allen, 2021). We consider the weighted Gower distance between observation $i$ and $j$ defined as $(1-\alpha)g(X_i, X_j) + \alpha g(y_i, y_j)$. Here, $g(X_i, X_j) = \sum_{l=1}^{p} g_{i,l} / p$, where $g_{i,l} = |X_{il} - X_{jl}|$ refers to the Gower distance between observation $i$ and $j$ for the $l$th feature and $R_l = \max_{i,j} |X_{il} - X_{jl}|$ is the range of the $l$th feature. Similarly, for continuous or count-valued $y$, the Gower distance is a range-normalized Manhattan distance, that is, $g(y_i, y_j) = \frac{|y_i - y_j|}{\max_{i,j} |y_i - y_j|}$. For categorical $y$, the distance $g(y_i, y_j)$ equals 0 if the categories match and 1 otherwise. Hence, the Gower distance is an average of scaled distances measured on all variables, where the distance calculated on each variable is scaled to range from 0 to 1 (maximum distance). It is appealing as all variables are converted to the same scale.

Denote $\nu^k_{ij}$ as the indicator which equals 1 if observation $j$ is among observation $i$’s $k$ nearest neighbors or vice versa, and 0 otherwise. Let $\alpha$ be a nonnegative tuning parameter between 0 and 1. Then, our recommendation for weights is given by the following:

\[
\omega_{ij} = \nu^k_{ij} \exp \left[ -\phi \left( (1-\alpha)g(X_i, X_j) + \alpha g(y_i, y_j) \right) \right].
\]
The $k$ nearest neighbors $y_{ij}^k$ are obtained based on the weighted Gower distance in the exponential. The tuning parameter $\alpha$ controls the level of supervision $y$ gives to data $X$. A larger $\alpha$ suggests putting more weights on the supervising auxiliary variables and hence favors the clustering of $y$. In practice, we suggest choosing $\alpha = \frac{D_y}{D_y + \|X - \bar{X}\|^2}$, where $D_y$ is the null deviance of supervising auxiliary variable $y$, and hence $\alpha$ is the ratio of null deviances between two sources. If the clustering signal in the supervising auxiliary variable is weak, our choice of weight will down-weight this variable and vice versa. This weighting scheme balances the contribution of $X$ and $y$. Yet, one may choose other weighting schemes based on the level of confidence in the supervising auxiliary variable as well. In addition to the recommended scheme for $\alpha$, motivated by the procedures mentioned above for choosing $\lambda$, we also provide two data-driven approaches to choose the parameters $\alpha$ and $\lambda$ based on the stability selection and information criterion in Web Appendix F if the user wants to tune these parameters.

In the presence of additional covariates, we can no longer calculate weights based on the distance between $y_i$ and $y_j$ as the supervising auxiliary variable $y$ now contains the effect of additional covariates $Z$. Recall the fusion penalty measures the dissimilarity between cluster centroids. To remove the effect of additional covariates in calculating weights, we suggest (i) first estimating the effect of covariates $\hat{\beta}$ by fitting SCC with weights not adjusted for covariates as usual and (ii) then calculating weights based on distances between the supervising auxiliary variable centroids $\hat{\theta}$ by removing the effect of covariates from the supervising auxiliary variable. This gives our adaptive SCC with covariate-adjusted weights, as detailed in Algorithm 2. Our adaptive SCC is similar to many adaptive approaches in the literature (Zou, 2006).

To remove the effects of additional covariates on the supervising auxiliary variable, we use the property of the link function: $g(E[y|Z]) = \theta + Z\tilde{\beta}$. A good estimate of $\theta$ would be removing the effect of covariate from the link function of $y$, that is, $\hat{\theta} = g(y) - Z\hat{\beta}$. Hence, we can get an estimate for the supervising auxiliary variable without the effect of additional covariates using the inverse link function: $\hat{y} = g^{-1}(\hat{\theta})$.

### 3 Simulation Studies

In this section, we evaluate the performance of our SCC methods and compare them with existing methods. For all the simulations, we design challenging scenarios where clustering either $X$ or $y$ alone cannot lead to good clustering results. We compare our SCC method with convex clustering, K-means, and hierarchical clustering using different distance metrics (Euclidean and Gower distances). Note that there are several linkage options for hierarchical clustering and we only report the linkage with the best clustering performance. We use the adjusted Rand index (Hubert & Arabie, 1985) to evaluate the accuracy of clustering results. The adjusted Rand index is a commonly used metric to measure the agreement between the estimated cluster label and the true underlying label. A larger adjusted Rand index (close to 1) indicates a good resemblance between the estimated and true labels. For the simulation results in the main paper, we use the proposed data-driven weight scheme of $\alpha$ in Section 2.5.2 based on the ratio of null deviances between two sources. In addition, for all methods, we assume that the oracle number of clusters is known for fair comparisons. Results when the number of clusters and level of supervision $\alpha$ are not fixed but estimated based on the data using tuning parameter selection approaches, are provided in Table 3 of the Web Appendix.

We first consider the base simulation where the supervising auxiliary variable $y$ is generated from the cluster centroid directly without additional covariates. In the base simulation, we study two designs of unlabeled data in which two different scenarios are considered. Specifically, the unlabeled data consists of $n = 120$ observations and $p = 30$ features with three clusters. Each cluster has an equal number of observations. The data are generated from the following model: $X_i \sim N(\mu_k, \sigma^2 I_p)$, where $i \in G_k$, $k = 1, 2, 3$ ($G_k$ refers to the observation
TABLE 1  Comparisons of the adjusted Rand index for SCC and existing methods; base simulation for Gaussian, binary, categorical, count-valued, and censored survival supervising auxiliary variables.

|        | Gaussian                | Binary              |
|--------|-------------------------|---------------------|
|        | S1 S2 H1 H2             | S1 H1 AS1           |
| Hclust on X | 0.71 (2.2e-2) 0.64 (3.2e-2) 0.20 (2.5e-2) 0.19 (7.4e-3) | 0.69 (1.2e-2) 0.34 (9.4e-2) 0.48 (2.3e-2) |
| Hclust on y | 0.35 (1.4e-2) 0.85 (4.0e-2) 0.58 (2.2e-2) 0.91 (1.3e-2) | 0.31 (6.8e-4) 0.35 (1.3e-3) 0.60 (3.1e-2) |
| Hclust on [X y] | 0.93 (1.4e-2) 0.87 (2.4e-2) 0.58 (1.1e-1) 0.39 (2.3e-2) | 0.74 (1.7e-2) 0.40 (8.7e-2) 0.50 (3.9e-2) |
| Hclust on [X y] Gower | 0.82 (4.5e-2) 0.91 (1.9e-2) 0.55 (2.2e-2) 0.22 (1.1e-2) | 0.84 (2.1e-3) 0.51 (5.6e-2) 0.92 (5.5e-2) |
| K-means on X | 0.66 (1.9e-2) 0.80 (2.2e-2) 0.20 (5.1e-3) 0.17 (5.2e-3) | 0.71 (2.2e-2) 0.22 (7.1e-3) 0.65 (2.8e-2) |
| K-means on [X y] | 0.86 (2.2e-2) 0.91 (1.3e-2) 0.27 (7.8e-3) 0.31 (9.0e-3) | 0.70 (1.7e-2) 0.21 (5.0e-3) 0.69 (3.0e-2) |
| Convex Clustering on X | 0.56 (1.5e-3) 0.21 (6.3e-2) 0.62 (2.2e-2) 0.13 (1.4e-2) | 0.56 (1.5e-3) 0.66 (1.1e-2) 0.32 (7.0e-2) |
| Convex Clustering on [X y] | 0.86 (6.6e-2) 0.58 (4.6e-2) 0.99 (3.3e-3) 0.60 (6.5e-2) | 0.56 (1.5e-3) 0.66 (1.2e-3) 0.32 (7.0e-2) |
| SCC     | 0.96 (1.1e-2) 0.95 (9.5e-3) 1.00 (3.3e-3) 0.97 (1.0e-2) | 0.85 (4.6e-3) 0.80 (3.3e-2) 0.96 (8.5e-3) |

|        | Categorical             |                    |
|--------|-------------------------|--------------------|
|        | S1 S2 H1 H2             | AS1 AS2            |
| Hclust on X | 0.71 (2.3e-2) 0.63 (2.0e-2) 0.34 (9.1e-2) 0.18 (1.1e-2) | 0.51 (1.1e-2) 0.51 (7.5e-3) |
| Hclust on y | 0.32 (5.2e-3) 0.81 (1.6e-2) 0.43 (3.8e-3) 0.81 (1.5e-2) | 0.80 (2.4e-2) 0.69 (1.9e-3) |
| Hclust on [X y] | 0.73 (2.5e-2) 0.66 (2.3e-2) 0.34 (9.1e-2) 0.18 (1.2e-2) | 0.51 (8.1e-3) 0.52 (7.1e-3) |
| Hclust on [X y] Gower | 0.81 (4.5e-3) 0.83 (1.8e-2) 0.68 (7.1e-2) 0.72 (8.0e-2) | 0.35 (2.8e-3) 0.79 (6.5e-2) |
| K-means on X | 0.74 (1.8e-2) 0.75 (2.0e-2) 0.21 (5.8e-3) 0.16 (6.4e-3) | 0.47 (1.4e-3) 0.84 (1.9e-2) |
| K-means on [X y] | 0.73 (1.7e-2) 0.75 (1.3e-2) 0.22 (7.9e-3) 0.16 (6.5e-3) | 0.48 (1.1e-2) 0.86 (2.3e-2) |
| Convex Clustering on X | 0.56 (1.2e-3) 0.14 (5.7e-2) 0.62 (2.4e-2) 0.15 (1.1e-2) | 0.41 (3.3e-2) 0.78 (7.1e-2) |
| Convex Clustering on [X y] | 0.56 (9.2e-4) 0.22 (7.5e-2) 0.65 (1.5e-2) 0.15 (1.5e-2) | 0.39 (3.2e-2) 0.82 (7.1e-2) |
| SCC     | 0.86 (2.0e-2) 0.89 (1.2e-2) 0.95 (3.3e-3) 0.85 (1.5e-2) | 0.87 (2.2e-2) 1.00 (0.0e-0) |

indices belonging to group $k$). The supervising auxiliary variable, $y_i$, is generated from different distributions with parameter $\mu_k$ based on the data type; the two sources have the shared group label which means $y_i \sim \phi(\mu_k)$, where $i \in G_k$ and $\phi$ is a distribution function. We denote $X_{G_k}$ and $y_{G_k}$ as the data points and their corresponding supervising auxiliary variable that belong to group $k$.

We consider two designs of the unlabeled data $X$: spherical (S) and half-moon (H). In terms of the half-moon data, we consider the standard simulated data of three interlocking half-moons as suggested by Chi and Lange (2015) and Wang and Allen (2021). For each design, we consider two scenarios where none of the data sources leads to perfect clustering results. In the first scenario (S1 and H1), $X_{G_1}$ and $X_{G_2}$ overlap while $X_{G_3}$ are separate from $X_{G_1}$ and $X_{G_2}$; $y_{G_1}$ and $y_{G_2}$ have two separate clusters, while $y_{G_3}$ are noisy and overlap with $y_{G_1}$ and $y_{G_2}$. In the second scenario (S2 and H2), $X_{G_1}$, $X_{G_2}$, and $X_{G_3}$ all overlap; $y$ has three separate clusters with little overlapping. Yet, $y$ is noisy and one cannot get perfect results by clustering $y$ alone. Hence, for each of the four simulation scenarios above, we create a challenging problem where good clustering results cannot be achieved by clustering either $X$ or
ComparisonsoftheadjustedRandindexforSCCandexistingmethods;supervisingauxiliaryvariableaffectedbyadditional
covariates;theunlabeleddataaresimulatedfromthesamesetupasS1inthebasesimulation,butthestrainingauxiliaryvariablesare
simulatedfromdifferentcentroidsaffectedbycovariates.WesaralsoleducresultsofK-meansclusteringontheconcatenatedmatrixofthe
unlabeleddataX and 𝜖,where 𝜖aretheresidualsofregressing y ontheadditionalcovariates Z.

|                        | Gaussian | Binary | Categorical | Count | Survival |
|------------------------|----------|--------|-------------|-------|----------|
| Hclust on X            | 0.68(2.1e-2) | 0.73(3.4e-2) | 0.70(2.8e-2) | 0.69(2.6e-2) | 0.73(2.1e-2) |
| Hclust on y            | 0.04(1.0e-2) | 0.22(2.6e-2) | 0.17(1.8e-2) | 0.07(7.9e-3) | 0.05(1.8e-2) |
| Hclust on [X y]        | 0.23(5.8e-2) | 0.76(3.0e-2) | 0.70(2.0e-2) | 0.56(2.7e-2) | 0.74(2.0e-2) |
| Hclust on [X y] Gower  | 0.57(2.8e-2) | 0.75(2.9e-2) | 0.70(2.3e-2) | 0.66(5.2e-2) | 0.59(1.5e-2) |
| K-means on X           | 0.72(2.6e-2) | 0.71(1.6e-2) | 0.68(1.5e-2) | 0.70(1.4e-2) | 0.69(2.6e-2) |
| K-means on [X y]       | 0.17(2.9e-2) | 0.75(2.0e-2) | 0.69(1.8e-2) | 0.49(2.7e-2) | 0.66(1.6e-2) |
| Convex Clustering on X | 0.56(1.2e-3) | 0.56(1.5e-3) | 0.56(3.0e-3) | 0.56(1.5e-3) | 0.56(1.4e-3) |
| Convex Clustering on [X y] | 0.40(6.4e-2) | 0.56(1.5e-3) | 0.55(4.9e-3) | 0.58(1.2e-2) | 0.56(1.4e-3) |
| K-means on [X 𝜖]      | 0.91(1.5e-2) | 0.77(2.2e-2) | 0.75(2.4e-2) | 0.79(2.4e-2) | 0.79(2.4e-2) |
| Adaptive SCC           | 0.99 (1.4e-2) | 0.85 (2.7e-2) | 0.94 (1.1e-2) | 0.89 (2.2e-2) | 0.83 (3.2e-2) |

We also consider additional simulations AS1 and AS2 with simulation details given in Web Appendix B.

Table 1 suggests that our SCC outperforms existing methods for different types of supervising auxiliary variables by leveraging information from both sources. For hierarchical clustering on spherical data, different distance metrics might perform comparably well on different types of supervising auxiliary variables. For example, hierarchical clustering with Euclidean distances works well for Gaussian and count-valued supervising auxiliary variables while hierarchical clustering with Gower distances works well for binary and categorical supervising auxiliary variables. Yet, our SCC performs comparatively well in terms of the best hierarchical clustering method for all cases. For nonspherical data, our method performs significantly better than hierarchical clustering. In addition, we consider different setups from the base simulation including varying numbers of features and unequal group sizes of 𝑛and verify that our method could still perform well in Web Appendix C.

Next, in the following simulation, we examine the performance of our SCC when the supervising auxiliary variable is affected by additional covariates. We use our adaptive SCC approach proposed in Section 2.5.2 to adjust for these additional covariates. The data X is generated from the same distribution as in scenario 1 (S1) of the base simulation described above. Still, we consider different types of supervising auxiliary variables. Yet, the supervising auxiliary variable 𝑦is now simulated from 𝑦𝑖∼ 𝜓(𝜇𝑘+ 𝑖′ 𝛽), where 𝑖∈ ℜ10 ∼ 𝑁(0, I10) and 𝛽𝑖∼ 𝑁(±3, 1); 𝜇𝑘is generated similarly in scenario 1 of the base simulation. We set the number of features for the additional covariates to be 10. Table 2 shows that our adaptive SCC performs the best by removing the effects of additional covariates from the supervising auxiliary variable; hence, our method clusters objects based on the exact centroids which form the groups. On the other hand, existing methods all do poorly as they perform clustering based on the supervising auxiliary variable y which is affected by additional covariates.

Finally, we evaluate the performance of our supervised convex biclustering. Note that, by the details of the simulation setup in Web Appendix B, the features are generated from two groups. We seek to cluster the observations and features simultaneously using supervised convex biclustering and evaluate the clustering accuracy of both subjects and features. Table 3 suggests that our supervised convex biclustering not only clusters common subjects but also groups similar features simultaneously and identifies checkerboard-like patterns.

Overall, we demonstrate the strong empirical performance of our SCC method, which includes the information from both sources to get better clustering results.

4 | CASE STUDY: DISCOVERING NEW SUBTYPES OF ALZHEIMER’S DISEASE

An important application of our proposed method is in clinical genomics, where the objective is to elucidate the
TABLE 3 Comparisons of the adjusted Rand index for the supervised convex biclustering and existing methods; the unlabeled data and supervising auxiliary variable are simulated from the same setup as the base simulation without additional covariates (Table 1) for the Gaussian and categorical supervising auxiliary variable. We compare the clustering accuracy of both observations and features for the biclustering algorithm. In addition, we examine the performance of the supervised convex biclustering method when the supervising auxiliary variable is affected by additional covariates in Web Appendix C.

|                  | Gaussian | Categorical |                  |                  |                  |                  |                  |
|------------------|----------|-------------|------------------|------------------|------------------|------------------|------------------|
|                  | S1       | S2          | H1               | H2               |                  |                  |                  |
|                  | ARI on Obs. | ARI on Feature | ARI on Obs. | ARI on Feature | ARI on Obs. | ARI on Feature | ARI on Obs. | ARI on Feature |
| Biclustering on X| 0.71 (2.4e-2) | 1.00 (0.0e-0) | 0.65 (3.7e-2) | 0.99 (1.3e-2) | 0.20 (2.5e-2) | 1.00 (0.0e-0) | 0.18 (1.4e-2) | 1.00 (0.0e-0) |
| Biclustering on [X y] | 0.94 (1.4e-2) | 1.00 (0.0e-0) | 0.88 (2.2e-2) | 0.99 (1.3e-2) | 0.58 (1.1e-1) | 1.00 (0.0e-0) | 0.38 (2.1e-2) | 1.00 (0.0e-0) |
| Convex Biclustering on X | 0.67 (5.1e-2) | 1.00 (0.0e-0) | 0.30 (6.0e-2) | 0.95 (4.8e-2) | 0.54 (4.2e-2) | 1.00 (0.0e-0) | 0.16 (1.3e-2) | 1.00 (0.0e-0) |
| Convex Biclustering on [X y] | 0.90 (5.8e-2) | 1.00 (0.0e-0) | 0.62 (4.6e-2) | 0.95 (4.8e-2) | 0.62 (5.2e-2) | 1.00 (0.0e-0) | 0.61 (6.4e-2) | 1.00 (0.0e-0) |
| Supervised Convex Biclustering | 0.96 (1.1e-2) | 1.00 (0.0e-0) | 0.95 (9.5e-3) | 1.00 (0.0e-0) | 1.00 (3.3e-3) | 1.00 (0.0e-0) | 0.97 (1.0e-2) | 1.00 (0.0e-0) |

For our analysis, to start with, we take the clinically measured global cognition score as the noisy supervising auxiliary variable. Global cognition score, as a summary measure of cognition proximal to death, is computed by averaging nineteen clinical cognitive tests conducted during a subject’s last clinical visit (Bennett et al., 2018). A higher value for the global cognition score indicates relatively healthier cognitive abilities. The ROSMAP data of older adults and little is known about the underlying genetic mechanisms that cause AD apart from a handful of genes. In this case study, we apply our SCC method to find a biologically meaningful group structure among both subjects and genomic profiles for AD by jointly analyzing RNA-Seq data obtained from the Religious Orders Study Memory and Aging Project (ROSMAP) Study (Bennett et al., 2018).
The resulting heatmap of RNA-Seq profiles is displayed in Figure 1A, where we order the subjects (rows) and genes (columns) according to the cluster assignment estimated by our SCC-biclustering with black dashed lines indicating cluster boundaries. In addition, the corresponding global cognition score is displayed on the left side of the heatmap. For comparison, we also show a completely supervised approach that treats the global cognition score as the true response variable. The heatmap generated by this completely supervised approach is shown in Figure 1B, where the subjects are ordered by the ascending global cognition score (from top to bottom) while the genes are ordered in ascending order of $p$-values obtained from univariate association tests of each gene with the global cognition score after adjusting for age. Finally, in Figure 1C, subjects and genes are ordered according to the cluster heatmap obtained from completely unsupervised hierarchical biclustering on unlabeled RNA-Seq data with the Euclidean distance metric and Ward linkage, as shown in Figure 1C.

Our results reveal, as shown in Figure 1A, the top cluster above the horizontal dashed line consists mostly of subjects with relatively low cognition scores whereas the bottom cluster is made up of subjects with generally higher cognition scores. For simplicity, we call the top cluster (bottom cluster) in Figure 1A the “low cognition cluster” (“high cognition cluster”) thereafter. A quick examination of RNA-Seq gene signatures across the low and high cognition clusters obtained from our SCC reveals clear differences in expression levels of many RNA-Seq genes, indicating possible genetic biomarkers responsible for influencing cognitive decline and onset of AD. On the other hand, even though completely unsupervised clustering of RNA-Seq data results in a heatmap (Figure 1C) with seemingly distinct genetic profiles across clusters, these estimated clusters are much less scientifically interpretable as each cluster contains subjects across the entire spectrum of global cognition score. Furthermore, the heatmap produced by the fully supervised approach in Figure 1B lacks almost any distinguishable patterns that
might reveal genetic differences between subjects with higher cognition and subjects with lower cognition. To briefly summarize, our SCC method leverages information from both the unlabeled RNA-Seq gene expression data and noisy supervising auxiliary variable, global cognition score, to recover more interpretable clusters of subjects and gene signatures, in contrast to either fully supervised or unsupervised methods which only estimates clusters using one data source. We further study the scientific validity of the clusters identified in Figure 1A and conclude that SCC discovers more scientifically interpretable cluster structures that shed unique light on AD; details are given in Web Appendix E. In particular, clusters from SCC reveal a collection of differentially expressed genes that have been validated in the AD literature, as well as new candidate genes for future studies of the genetic basis for AD. In addition, we also examine the heterogeneity among subjects uncovered by SCC, identifying atypical subjects that might possess certain degrees of so-called cognitive resilience (CR).

5 | DISCUSSION

In this paper, we develop a novel SCC method that leverages the information from both supervising auxiliary variables and unlabeled data. Our method, in contrast to existing semi-supervised clustering approaches, is the first one to directly use some outcomes of interest to help cluster unlabeled data. In particular, our SCC borrows strength from both information sources and yields more scientifically interpretable group structures that may be hidden in completely unsupervised analyses of data.

One question that is worth further investigating is, whether we should use SCC, and, how practitioners can tell whether the supervising auxiliary variable is useful for finding group structures. Further research could investigate when to apply SCC and how much supervision is warranted for given problems.

We apply our method to a high-dimensional genomics case study. Yet, our approach may find applications in a variety of fields such as electronic health records, online market segmentation, and text mining, among the many other clustering applications. Overall, we develop a novel, unified approach to an interesting but challenging problem that leads to more scientifically interpretable clustering results and opens many avenues for future research.

ACKNOWLEDGMENTS

The authors thank the editor, the associate editor, and two anonymous referees for their valuable comments and suggestions. The authors acknowledge support from NSF DMS-1554821, NSF NeuroNex-1707400, NIH R01GM140468, and NSF DMS-2210837, and thank Dr. Joshua Shulman for discussions on the ROSMAP data. The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Study data were provided by the Rush Alzheimer’s Disease Center, Rush University Medical Center, Chicago, Illinois. Data collection was supported through funding by NIA grants R01AG15819 and R01AG36836.

DATA AVAILABILITY STATEMENT

The data that support the findings in this paper are available from the AD Knowledge Portal. Restrictions apply to the availability of these data, which were used under license in this paper. Data are available at https://adknowledgeportal.org with the permission of the AD Knowledge Portal.

ORCID

Minjie Wang © https://orcid.org/0000-0002-6307-0028

REFERENCES

Ali, B.B. & Massmoudi, Y. (2013) K-means clustering based on Gower similarity coefficient: A comparative study. In 2013 5th International Conference on Modeling, Simulation and Applied Optimization (ICMSAO). Piscataway, NJ: IEEE, pp. 1–5.

Bair, E. & Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. PLoS Biology, 2(4), e108.

Basu, S., Banerjee, A. & Mooney, R. (2002) Semi-supervised clustering by seeding. In Proceedings of the 9th International Conference on Machine Learning (ICML-2002). San Francisco, CA: Morgan Kaufmann, pp. 27–34.

Basu, S., Banerjee, A. & Mooney, R. (2004) Active semi-supervision for pairwise constrained clustering. In Proceedings of the 2004 SIAM International Conference on Data Mining. Philadelphia, PA: SIAM, pp. 333–344.

Bennett, D., Buchman, A., Boyle, P., Barnes, L., Wilson, R. & Schneider, J. (2018) Religious orders study and rush memory and aging project. Journal of Alzheimer’s Disease, 64(s1), 161–189.

Chi, E.C., Allen, G.I. & Baraniuk, R.G. (2017) Convex biclustering. Biometrics, 73(1), 10–19.

Chi, E.C. & Lange, K. (2015) Splitting methods for convex clustering. Journal of Computational and Graphical Statistics, 24(4), 994–1013.

Deng, W., Lai, M.-J., Peng, Z. & Yin, W. (2017) Parallel multi-block ADMM with o(1/k) convergence. Journal of Scientific Computing, 71(2), 712–736.

Fang, Y. & Wang, J. (2012) Selection of the number of clusters via the bootstrap method. Computational Statistics & Data Analysis, 56(3), 468–477.

Gower, J.C. (1971) A general coefficient of similarity and some of its properties. Biometrics, 27(4), 857–871.

Hocking, T.D., Joulin, A., Bach, F. & Vert, J.-P. (2011) Clusterpath: an algorithm for clustering using convex fusion penalties. In Proceedings of the 28th International Conference on Machine Learning (ICML-2011). New York: ACM, pp. 745–752.
Hubert, L. & Arabie, P. (1985) Comparing partitions. *Journal of Classification*, 2(1), 193–218.

Hummel, M., Edelmann, D. & Kopp-Schneider, A. (2017) Clustering of samples and variables with mixed-type data. *PLoS ONE*, 12(11), e0188274.

Johnson, W., Li, C. & Rabinovic, A. (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. *Bioinformatics*, 8(1), 118–127.

Koestler, D.C., Marsit, C.J., Christensen, B.C., Karagas, M.R., Bueno, R., Sugarbaker, D.J., Kelsey, K.T. & Houseman, E.A. (2010) Semi-supervised recursively partitioned mixture models for identifying cancer subtypes. *Bioinformatics*, 26(20), 2578–2585.

Lin, T., Ma, S. & Zhang, S. (2015) On the global linear convergence of the ADMM with multiblock variables. *SIAM Journal on Optimization*, 25(3), 1478–1497.

Lindsten, F., Ohlsson, H. & Ljung, L. (2011) *Just relax and come clustering!*: A convexification of k-means clustering. Linköping, Sweden: Linköping University Electronic Press.

Ma, S., Xue, L. & Zou, H. (2013) Alternating direction methods for latent variable Gaussian graphical model selection. *Neural Computation*, 25(8), 2172–2198.

Pelckmans, K., de Brabanter, J., Suykens, J. & de Moor, B. (2005) Convex clustering shrinkage. In *PASCAL Workshop on Statistics and Optimization of Clustering*.

Radchenko, P. & Mukherjee, G. (2017) Convex clustering via $\ell_1$ fusion penalization. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 79(5), 1527–1546.

Tan, K.M. & Witten, D. (2015) Statistical properties of convex clustering. *Electronic Journal of Statistics*, 9(2), 2324–2347.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. & Knight, K. (2005) Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1), 91–108.

Wang, B., Zhang, Y., Sun, W.W. & Fang, Y. (2018) Sparse convex clustering. *Journal of Computational and Graphical Statistics*, 27(2), 393–403.

Wang, J. (2010) Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4), 893–904.

Wang, M. & Allen, G.I. (2021) Integrative generalized convex clustering optimization and feature selection for mixed multi-view data. *Journal of Machine Learning Research*, 22, 1–73.

Weylandt, M., Nagorski, J. & Allen, G.I. (2020) Dynamic visualization and fast computation for convex clustering via algorithmic regularization. *Journal of Computational and Graphical Statistics*, 29(1), 87–96.

Xing, E., Jordan, M., Russell, S.J. & Ng, A. (2002) Distance metric learning with application to clustering with side-information. In *Advances in neural information processing systems*, volume 15. Cambridge, MA: MIT Press, pp. 521–528.

Zhu, J. & Hastie, T. (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics*, 5(3), 427–443.

Zhu, Y. (2017) An augmented ADMM algorithm with application to the generalized Lasso problem. *Journal of Computational and Graphical Statistics*, 26(1), 195–204.

Zou, H. (2006) The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), 1418–1429.

**SUPPORTING INFORMATION**

Web Appendices referenced in Sections 2, 3, and 4, as well as the MATLAB codes are available with this paper at the Biometrics website on Wiley Online Library.

**How to cite this article:** Wang, M., Yao, T. & Allen, G.I. (2023) Supervised convex clustering. *Biometrics*, 79, 3846–3858. [https://doi.org/10.1111/biom.13860](https://doi.org/10.1111/biom.13860)