mathematics

MDPI

Article

Expectile Sparse Clustering for Mixed-Type Data with Data-Driven Ordinal Scoring

Ismail Arjdal [1,*], Mohamed Alahiane [2], Echarif Elharfaoui [1] and Mustapha Rachdi [3]

[1] Department of Mathematics, Faculty of Sciences, Chouaib Doukkali University, El Jadida 24000, Morocco; elharfaoui.e@ucd.ac.ma
[2] LERSEM Laboratory, National School of Business and Management, Chouaib Doukkali University, Corner Avenue Ahmed Chaouki and Rue de Fès, BP.122, El Jadida 24000, Morocco; alahiane.mohamed@ucd.ac.ma
[3] AGEIS Laboratory, Grenoble Alpes University (UGA), UFR SHS, BP.47, CEDEX 09, 38040 Grenoble, France; mustapha.rachdi@univ-grenoble-alpes.fr
* Correspondence: arjdal.i537@ucd.ac.ma

Abstract

High-dimensional datasets combining continuous, ordinal, and nominal variables are increasingly prevalent in modern applications such as marketing analytics, biostatistics, and social sciences. Clustering in such settings is challenging because continuous variables may exhibit asymmetric distributional structure, ordinal variables have a natural ordering but no canonical numerical spacing, and irrelevant variables can obscure the underlying cluster structure. We propose ExSC-$\tau\theta$, an adaptive sparse clustering framework for high-dimensional mixed-type data that addresses these three issues within a unified variable-selection framework. For continuous variables and Azzalini-scored ordinal variables, we introduce an expectile-based asymmetric dissimilarity parameterized by $\tau \in (0,1)$, which reduces to the squared Euclidean dissimilarity at $\tau = \frac{1}{2}$ and allows for directional differences to receive unequal quadratic penalties. For ordinal variables, we extend Azzalini's score transformation to the unsupervised setting by estimating the transformation parameters from a cluster-quality criterion, thereby obtaining data-driven non-equidistant numerical representations. For nominal variables, we employ a co-occurrence-based dissimilarity that incorporates dependence among categorical variables. The resulting type-specific dissimilarities are integrated through separate ℓ_1-constrained variable-weight vectors and a variable-wise adjusted Davies–Bouldin objective. We establish screening consistency under explicit signal-strength, sparsity, dimensionality, and regularity conditions. Simulation studies covering different numbers of clusters and distributional heterogeneity demonstrate that the proposed framework is particularly effective for ordinal-variable recovery and for clustering structures in which symmetric dissimilarities are inadequate. Two real data applications, Steel Annealing and Credit Approval, illustrate the method, including the interpretability of the learned ordinal representation, the variable selection and the adaptive asymmetry parameter.

Keywords: mixed-type clustering; sparse clustering; expectile dissimilarity; ordinal variables; Azzalini score; variable selection; high-dimensional data

MSC: 62H30; 62H17; 62J07

Check for updates

Academic Editor: Heng Lian

Received: 29 September 2026
Revised: 26 October 2026
Accepted: 27 October 2026
Published: 30 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Large-scale heterogeneous datasets are now ubiquitous across domains such as marketing analytics, risk modeling, medical registries, and socio-economic surveys. A common

feature of these datasets is the co-occurrence of multiple variable types: continuous measurements such as age or income, ordinal ratings such as education level or satisfaction score, and nominal categories such as profession or region. When the number of variables p is large relative to the sample size n, unsupervised discovery of group structure in such data presents several compounding difficulties: the variable types call for different notions of dissimilarity, ordinal levels have an order but no canonical numerical spacing, continuous variables may be asymmetrically distributed, and irrelevant variables can mask the cluster structure.

Classical clustering methods such as k-means, hierarchical clustering, and Gaussian mixture models rely on the Euclidean metric or a Gaussian likelihood, both of which assume continuous, roughly symmetric, and homogeneous data. In high-dimensional mixed-type settings, these assumptions are violated on every count: Euclidean distances become concentrated and uninformative as the dimension grows [1], ordinal and nominal variables have no natural Euclidean representation, and the absence of variable selection allows for noise variables to mask the true cluster structure.

1.1. Mixed-Type Clustering

A large literature addresses heterogeneous variables. Classical tools include k-prototypes [2], Gower's dissimilarity [3], and model-based approaches including VarSelLCM [4]. Recent work follows two main lines. Distance- and partition-based methods modify or learn the dissimilarity: a modified and weighted Gower distance [5], a learned unified distance metric for heterogeneous attributes [6], weighted partition-based clustering with dynamic categorical weighting [7], and mixed fuzzy C-means [8]. Model-based methods extend mixture models to mixed-type data: composite-likelihood methods for parsimonious model-based clustering [9] and mixture models for skewed mixed-type data [10]. Comparative studies confirm that the performance of these approaches depends strongly on the data and on the treatment of each variable type [11]. However, these methods weight or model the variables but, to our knowledge, none of them performs sparse variable selection, that is, sets the weight of uninformative variables exactly to zero, which is essential when p is large.

1.2. Ordinal Variables

Ordinal variables deserve specific attention. Szepannek et al. [12] show empirically that naive ordinal encodings can severely distort the cluster structure, and propose an extension of k-prototypes with ordinal-specific distances. Their method, like the other approaches above, keeps all variables and performs no variable selection, and ordinal encodings that are fixed beforehand or that do not depend on the cluster structure remain the rule in the literature. Azzalini [13] proposes a flexible score transformation that assigns non-equidistant numerical scores to ordinal levels, but in a supervised (regression) setting.

1.3. Sparse Clustering

Sparse clustering methods [14] address high dimensionality through ℓ_1-penalized feature weighting, assigning zero weight to variables that do not contribute to cluster separation. Extensions including the DBI-SC method of [15] improve variable-specific discrimination through a Davies–Bouldin-type criterion and accommodate mixed-type data: continuous variables enter through squared Euclidean differences, while ordinal and nominal variables are both processed through a co-occurrence distance, i.e., ordinal variables are treated as unordered categories. Consequently, DBI-SC does not exploit the ordering of ordinal levels, does not learn a numerical spacing for them, and uses a symmetric dissimilarity for continuous variables. Sparse k-means, in turn, handles ordinal and nominal

variables only through ad hoc preprocessing such as dummy encoding (which destroys ordinal structure) or arbitrary integer scoring (which imposes unwarranted spacing).

1.4. Quantile and Expectile Clustering

Quantile- and expectile-based clustering methods [16–18] provide dissimilarities that are sensitive to distributional shape and robust to outliers. These methods are, however, restricted to continuous variables and do not incorporate sparsity or mixed-type integration.

1.5. Research Gap and Contributions

Taken together, the existing methods either select variables but treat ordinal variables as unordered or as equally spaced (sparse methods), or handle mixed types and ordinal variables but keep all variables (recent distance-based and model-based methods, including [12]). This paper introduces **ExSC-$\tau\theta$**, a sparse clustering method that unifies robust metric dissimilarities, principled ordinal encoding, nominal co-occurrence distances, and variable selection within a single optimization framework. The method makes four specific contributions.

First, we replace the standard squared Euclidean distance with an expectile-based dissimilarity parameterized by $\tau \in (0, 1)$, applied uniformly to both continuous variables (through raw standardized differences) and Azzalini-scored ordinal variables (through score differences). At $\tau = \frac{1}{2}$, the standard distance is exactly recovered, making this a strict generalization of existing DBI-based sparse clustering. For $\tau \neq \frac{1}{2}$, the dissimilarity assigns asymmetric quadratic penalties to positive and negative gaps, capturing directional skewness in the cluster structure. The parameter τ is estimated from the data jointly with the clustering, with no user specification required.

Second, we extend the score encoding of [13] to the unsupervised setting by estimating the shape parameters θ_j from a cluster-quality criterion rather than fixing them externally. Unlike previous applications of Azzalini's method, where θ_j is selected by minimizing a supervised regression criterion, our two-stage estimation strategy maximizes the variable-wise adjusted Davies–Bouldin score at a provisional partition, producing ordinal scores whose spacing is tailored to the discovered cluster structure. Once scored, ordinal variables participate in the same expectile dissimilarity as continuous variables, governed by the shared parameter τ.

Third, continuous and ordinal dissimilarities (both of metric type, both governed by τ) are combined with the nominal co-occurrence dissimilarity into a single weighted pairwise dissimilarity, with separate ℓ_1-penalized weight vectors for metric and nominal variable groups. The optimization simultaneously assigns zero weight to uninformative variables and refines the cluster partition, without requiring the user to pre-specify variable importance.

Fourth, we establish the screening consistency of the ExSC-$\tau\theta$ estimator: under mild conditions on signal strength, dimensionality growth, and sparsity parameter calibration, all truly informative variables are included in the estimated support with probability tending to one. The proof accommodates subexponential concentration for metric variables and Bernstein-type concentration for nominal variables, and is robust to the plug-in of the pre-estimated ordinal parameters $\hat{\theta}_j$. This guarantee concerns variable selection; it does not cover recovery of the partition itself, whose accuracy is assessed empirically (see Sections 5 and 6).

To the best of our knowledge, ExSC-$\tau\theta$ is the first sparse clustering method that combines (i) variable selection for mixed-type data, (ii) data-driven, clustering-oriented numerical scoring of ordinal levels, and (iii) a learnable asymmetric (expectile) dissimilarity applied to both continuous and scored-ordinal variables.

1.6. Organization of the Paper

The remainder of the paper is organized as follows. Section 2 reviews the related literature and states the research gap. Section 3 presents the method: the treatment of each variable type, the expectile dissimilarity, the learnable ordinal scores, the nominal dissimilarity and the unified dissimilarity. Section 4 introduces the sparse objective, the algorithm and its complexity, and establishes the screening consistency of the estimator, with its scope. Section 5 reports the simulation studies (software, competitors, evaluation criteria, results, ablation, computing time). Section 6 applies the method to two real datasets, Steel Annealing and Credit Approval. Section 7 concludes.

2. Related Work

2.1. Sparse Clustering

Sparse clustering for continuous data was introduced by [14] through an ℓ_1-penalized framework that weights variables by their contribution to a between-cluster sum of squares. The DBI-SC method of [15] replaces the between-cluster sum of squares with a variable-wise Davies–Bouldin ratio, improving discrimination and providing screening consistency guarantees. Sparse k-means is restricted to continuous variables, and ordinal or nominal variables can only be incorporated through dummy encoding or integer coding, which destroys ordinal structure, inflates dimensionality or imposes arbitrary spacing. DBI-SC accommodates mixed-type data, but ordinal and nominal variables are both processed through a co-occurrence distance, so that ordinal variables are treated as unordered categories.

2.2. Mixed-Type Clustering

For mixed-type data, K-prototypes [2] combines k-means for continuous variables with k-modes for categorical variables, but requires a user-specified weight balancing the two parts. Gower's dissimilarity [3] provides a general-purpose measure for mixed types and is commonly used with PAM clustering [19], but offers no variable selection and depends on an additive structure that treats all variable types symmetrically. The VarSelLCM method [4] performs simultaneous clustering and variable selection within a latent class model for mixed data, but assumes conditional independence across variables given cluster membership and normal or multinomial distributions within clusters, assumptions that are frequently violated in practice. Marbac and Sedki [20] select variables in model-based clustering through the integrated complete-data likelihood, the criterion underlying VarSelLCM.

Several reviews survey clustering of mixed-type data: Foss et al. [21] compare distance metrics and clustering methods for mixed interval and categorical data and give guidelines for their use, and Ahmad and Khan [22] propose a taxonomy of mixed-data clustering into five research themes. Hennig and Liao [23] compare a latent class mixture model with dissimilarity-based methods such as k-medoids for metric, ordinal and nominal variables. KAMILA [24] balances continuous and nominal variables without dummy coding, and Mbuga and Tortora [25] extend spectral clustering to mixed-type data through type-specific dissimilarities combined by a weighted sum.

More recent contributions follow two lines. Distance- and partition-based methods modify or learn the dissimilarity or the attribute weights: Liu et al. [5] propose a modified and weighted Gower distance for mixed-type data; Zhang et al. [6] learn a unified distance metric for heterogeneous attributes; Dutt and Ali [7] propose a weighted partition-based method with dynamic weighting of categorical attributes; and a mixed fuzzy C-means algorithm is proposed in [8]. Model-based methods extend mixture models to mixed-type data: Ranalli and Rocci [9] use composite likelihood for parsimonious model-based clustering, and Alamer et al. [10] propose a mixture model for skewed mixed-type data. Ghattas and Sanchez-

San-Benito [11] compare several clustering approaches for mixed-type data. These methods weight or model the variables, but, to our knowledge, none of them performs sparse variable selection in the sense of setting the weights of uninformative variables exactly to zero.

2.3. Ordinal Variables

The treatment of ordinal variables in clustering has received comparatively little attention. Ref. [12] provides a systematic empirical study showing that naive integer encoding of ordinal levels degrades clustering performance substantially relative to methods that respect ordinal structure. The score transformation of [13] offers a principled approach: it maps ordinal levels to non-equally spaced numerical scores whose spacing is estimated from data by maximizing a regression fit criterion. Our method adapts this transformation to the unsupervised setting by replacing the regression criterion with a cluster-quality objective, and further extends it by applying an expectile dissimilarity to the resulting scores, integrating both elements within a sparse framework.

2.4. Quantile and Expectile Clustering

Robust and distribution-aware distances have also been explored. Quantile-based clustering [16] defines cluster membership through quantile regression, providing robustness to outliers and asymmetric distributions. K-expectile clustering [17] replaces the mean-based centroid with an expectile, assigning asymmetric quadratic penalties to deviations, with fuzzy extensions developed in [18]. These methods are restricted to continuous variables, do not incorporate variable selection, and have not been extended to mixed-type data. Our expectile dissimilarity is conceptually related but serves a different role: rather than replacing the cluster centroid, it modifies the pairwise dissimilarity used in a centroid-free sparse framework, applies uniformly to both continuous and ordinal variables, and its asymmetry parameter τ is estimated jointly with the clustering rather than pre-specified.

2.5. Positioning and Research Gap

Table 1 summarizes the key properties of competing methods. To our knowledge, none of the methods reviewed above combines ℓ_1-type sparse variable selection with a data-driven ordinal scoring and an asymmetric dissimilarity for continuous and ordinal variables: sparse methods select variables but treat ordinal variables as unordered or equally spaced, whereas the recent distance-based and model-based methods, including the ordinal-aware method of [12], handle mixed or ordinal data but keep all variables. ExSC-$\tau\theta$ is designed to fill this gap: among the methods of Table 1, it is the only one that handles mixed types, uses a learnable asymmetric metric dissimilarity, encodes ordinal variables in a data-driven, clustering-oriented way, and performs variable selection with a screening guarantee.

Table 1. Comparison of clustering methods across key properties. DBI-SC treats ordinal variables as nominal. ✓: supported; ◦: partial support or limited; ×: not supported.

Method	Mixed Types	Robust/ Asymmetric Metric Distance	Ordinal-Aware Encoding	Variable Selection	Theory
Sparse k-means [14]	×	×	×	✓	✓
DBI-SC [15]	✓	×	×	✓	✓
K-prototypes [2]	✓	×	×	×	×
PAM + Gower [3]	✓	×	◦	×	×
VarSelLCM [4]	✓	×	×	✓	◦
K-expectile [17]	×	✓	×	×	×
DAFI-Gower [5]	✓	×	◦	×	◦
Ordinal-aware K-prototypes [12]	✓	×	✓	×	✓

Table 1. *Cont.*

Method	Mixed Types	Robust/ Asymmetric Metric Distance	Ordinal-Aware Encoding	Variable Selection	Theory
Composite-likelihood mixed clustering [9]	✓	×	✓	∘	✓
Mixed fuzzy *C*-means [8]	✓	×	∘	×	∘
Skewed mixed-type mixture model [10]	✓	∘	∘	×	✓
Unified distance metric learning [6]	✓	×	✓	∘	∘
ExSC-$\tau\theta$ (this paper)	✓	✓	✓	✓	✓

3. Methodology

We consider a dataset of n independent observations to be partitioned into K clusters. Each observation $i \in \{1, \dots, n\}$ is described by a mixed-type vector

$$X_i = (x_{i1}, \dots, x_{i,p+r+q})^\top, \tag{1}$$

where the first p components are continuous, the next r are ordinal, and the final q are nominal. The full data matrix is $X = (X_1, \dots, X_n)^\top$. The domain of variable j is

$$\mathcal{X}_j = \begin{cases} \mathbb{R}, & 1 \le j \le p, \\ \{1, \dots, M_j\}, & p+1 \le j \le p+r, \\ \{1, \dots, L_j\}, & p+r+1 \le j \le p+r+q. \end{cases} \tag{2}$$

The goal is to cluster the n observations and simultaneously identify the variables that drive the cluster structure, while fully respecting the intrinsic nature of each variable type. We construct type-specific dissimilarities in Sections 3.2–3.4, combine them into a unified weighted pairwise dissimilarity in Section 3.5, and define the cluster dispersion criteria in Section 3.6. Section 3.1 first summarizes how each variable type is handled.

3.1. Handling of the Variable Types

Each variable type is mapped to a dissimilarity adapted to its nature, as summarized in Table 2. Continuous variables are standardized and compared through the expectile dissimilarity; ordinal variables are first transformed to real-valued scores whose spacing is learned from the data and are then compared through the same expectile dissimilarity; nominal (including binary) variables are compared through a co-occurrence dissimilarity. Count or discrete numeric variables are treated as continuous after standardization, or as ordinal when they take few ordered levels. In the k-means step of Algorithm 1, nominal variables are additionally dummy-coded (Section 4.6).

Table 2. Treatment of each variable type in ExSC-$\tau\theta$.

Type	Domain	Representation	Dissimilarity $d_{ii',j}$	Weight
Continuous	$\mathbb{R}$	standardized	$d_\tau(x_{ij} - x_{i'j})$, Equation (9)	w_j
Ordinal	$\{1, \dots, M_j\}$, ordered	Azzalini scores $q_{jk}(\hat{\theta}_j)$, Equations (12) and (13)	$d_\tau(q_{j,x_{ij}} - q_{j,x_{i'j}})$, Equation (10)	w_j
Nominal	$\{1, \dots, L_j\}$	integer-coded factor	co-occurrence $\tilde{\delta}$, Equation (20), in $[0,1]$	z_j
Binary	$\{1,2\}$	nominal with $L_j = 2$	Equation (20)	z_j
Count/discrete	$\mathbb{N}$	continuous (or ordinal if few levels)	Equation (9) (or (10))	w_j

Algorithm 1 ExSC-$\tau\theta$: Coordinate descent

Require: Data X, number of clusters K, sparsity bounds s_w, s_z.

Preprocessing. Standardize continuous variables to zero mean and unit variance.

Score estimation. Run preliminary k-means (5 iterations, equal weights, $\tau = \frac{1}{2}$) to obtain $C^{(0)}$. For each ordinal variable j, compute $\hat{\theta}_j$ by solving (17) via Nelder–Mead. Fix all scores $q_{jk}(\hat{\theta}_j)$.

Initialization. Set $w_j^{(0)} = (p+r)^{-1/2}$ for all $j \le p+r$, $z_j^{(0)} = q^{-1/2}$ for all $j > p+r$, and $\tau^{(0)} = \frac{1}{2}$.

repeat

  Step 1 *(Update C):* With $(\mathbf{w}, \mathbf{z}, \tau)$ fixed, obtain a new partition by weighted k-means (Lloyd's iterations on the $\sqrt{w_j}$- and $\sqrt{z_j}$-scaled representation; nominal variables dummy-coded).

  Step 2 *(Update $\mathbf{w}$):* With $(C, \mathbf{z}, \tau)$ fixed, for each $j \le p+r$, compute $a_j = \mathrm{adjDBI}_j^{(w)}(C, \tau, \hat{\boldsymbol{\theta}})$ and set $\mathbf{w} \leftarrow \mathcal{S}(\mathbf{a}^+, \lambda)/\|\mathcal{S}(\mathbf{a}^+, \lambda)\|_2$, where $\mathcal{S}(x, \lambda) = \mathrm{sign}(x)(|x| - \lambda)_+$ is the soft-threshold operator and $\lambda \ge 0$ is chosen so that $\|\mathbf{w}\|_1 = s_w$.

  Step 3 *(Update $\mathbf{z}$):* Proceed analogously to Step 2, using $a_j = \mathrm{adjDBI}_j^{(z)}(C)$ for $j > p+r$.

  Step 4 *(Update τ):* With $(C, \mathbf{w}, \mathbf{z})$ fixed, maximize the metric component of (28) over $\tau \in (0,1)$ by golden-section search.

until convergence: $\delta_w^{(t)} + \delta_z^{(t)} < 10^{-4}$, where $\delta_w^{(t)} = \|\mathbf{w}^{(t)} - \mathbf{w}^{(t-1)}\|_1 / \|\mathbf{w}^{(t-1)}\|_1$ and analogously $\delta_z^{(t)}$.

return $(\hat{C}, \hat{\mathbf{w}}, \hat{\mathbf{z}}, \hat{\tau}, \hat{\boldsymbol{\theta}})$.

3.2. Expectile Dissimilarity for Metric Variables

After the Azzalini score transformation of Section 3.3, both continuous and Azzalini-scored ordinal variables contribute real-valued, normalized pairwise differences. We therefore apply a single expectile dissimilarity to both variable types, governed by one global asymmetry parameter $\tau \in (0,1)$ that is estimated jointly with the clustering.

Definition 1 (Expectile loss)**.** *For $\tau \in (0,1)$, the expectile loss of $u \in \mathbb{R}$ is*

$$e_\tau(u) = \left|\tau - \mathbf{1}_{u<0}\right| u^2 = \begin{cases} \tau\, u^2, & u \ge 0, \\ (1-\tau)\, u^2, & u < 0. \end{cases} \tag{3}$$

The expectile loss equals $\frac{1}{2}u^2$ at $\tau = \frac{1}{2}$ and assigns different quadratic penalties to positive and negative deviations [17]. We define the associated pairwise expectile-based dissimilarity as

$$d_\tau(u) := e_\tau(u) + e_{1-\tau}(-u) = c(\tau, u)u^2, \tag{4}$$

where

$$c(\tau, u) = \begin{cases} 2\tau, & u > 0, \\ 2(1-\tau), & u \le 0. \end{cases} \tag{5}$$

Thus, $d_\tau(u) \ge 0$ for all u, and

$$d_\tau(u) \ne d_\tau(-u) \qquad \text{when } \tau \ne \tfrac{1}{2}. \tag{6}$$

Hence, d_τ should be interpreted as an asymmetric pairwise dissimilarity rather than as a metric or a symmetric distance. The pairwise form d_τ is introduced in this paper. At $\tau = \frac{1}{2}$,

$$d_{1/2}(u) = u^2, \tag{7}$$

so the conventional squared Euclidean contribution is recovered. Moreover, since $c(\tau, u) + c(\tau, -u) = 2$ for every $u \ne 0$,

$$d_\tau(u) + d_\tau(-u) = 2u^2 \qquad \text{for every } u \in \mathbb{R}. \tag{8}$$

Identity (8) is used repeatedly below: when a dissimilarity is summed over ordered pairs, both orders appear and the asymmetry cancels.

3.2.1. Continuous Variables

All continuous variables are standardized to zero mean and unit variance prior to analysis. The pairwise expectile dissimilarity for continuous variable $j \leq p$ between observations i and i' is

$$d^{(C)}_{ii',j}(\tau) = d_\tau(x_{ij} - x_{i'j}) = c(\tau, x_{ij} - x_{i'j}) \cdot (x_{ij} - x_{i'j})^2. \tag{9}$$

Setting $\tau = \frac{1}{2}$ yields $d^{(C)}_{ii',j}(\frac{1}{2}) = (x_{ij} - x_{i'j})^2$, recovering the standard squared Euclidean distance used in [15]. For $\tau \neq \frac{1}{2}$, the dissimilarity penalizes positive and negative gaps asymmetrically: $\hat{\tau} > \frac{1}{2}$ indicates that upward gaps are more influential for the cluster structure, a feature that may arise, for instance, when within-cluster distributions are right-skewed. The parameter τ is estimated from the data jointly with the clustering, as detailed in Section 4.1.

3.2.2. Ordinal Variables

Once ordinal variable $j \in \{p+1, \ldots, p+r\}$ has been transformed to normalized real-valued scores $q_{j,x_{ij}}(\hat{\theta}_j)$ (Section 3.3), its pairwise dissimilarity takes exactly the same expectile form:

$$d^{(O)}_{ii',j}(\tau, \hat{\theta}_j) = d_\tau\big(q_{j,x_{ij}}(\hat{\theta}_j) - q_{j,x_{i'j}}(\hat{\theta}_j)\big) = c\big(\tau, q_{j,x_{ij}}(\hat{\theta}_j) - q_{j,x_{i'j}}(\hat{\theta}_j)\big) \cdot \big(q_{j,x_{ij}}(\hat{\theta}_j) - q_{j,x_{i'j}}(\hat{\theta}_j)\big)^2. \tag{10}$$

The parameters θ_j and τ govern strictly different aspects of this dissimilarity. The shape parameter θ_j controls the absolute spacing between ordinal levels: how far apart levels k and $k+1$ are on the real line after transformation. The asymmetry parameter τ controls the directional penalty on pairwise score gaps in the clustering loss: whether a positive score difference is penalized more heavily than a negative one of equal magnitude. The two parameters are therefore not redundant. Even when $\theta_j = (0, 1)$ (standard normal scores, symmetric spacing), a value $\tau \neq \frac{1}{2}$ captures a directional asymmetry in the cluster structure that θ_j alone cannot represent. Conversely, θ_j shapes the score distribution independently of whether the loss is symmetric or not.

This shared τ creates a clean two-group structure in the unified distance: all metric variables, continuous and Azzalini-scored ordinal alike, share the same global asymmetry coefficient, while nominal variables form a separate group with their own co-occurrence-based scale (Section 3.4).

3.3. Ordinal Variables: Learnable Azzalini Scores

Ordinal variables carry a natural ordering of their levels but no canonical numerical spacing. Assigning integer codes $1, 2, \ldots, M_j$ imposes equal spacing that may distort distances, while one-hot encoding discards the ordering entirely. We adopt the score-based encoding of [13], extended here by making the transformation parameters estimable from the data within the clustering framework.

3.3.1. Parametric Score Transformation

Let X_j be an ordinal variable with ordered levels $\{1, \ldots, M_j\}$. Starting from equally spaced quantile seeds

$$u_k = \frac{k}{M_j + 1}, \quad k = 1, \ldots, M_j, \tag{11}$$

the transformed scores are defined as

$$q_{jk}(\theta_j) = Q(u_k; \theta_j), \tag{12}$$

where $Q(\cdot;\theta_j)$ is a parametric quantile function indexed by $\theta_j \in \Theta \subset \mathbb{R}^2$. Because the scores enter a dissimilarity that is invariant to affine transformations [13], we impose the normalization

$$\sum_{k=1}^{M_j} q_{jk}(\theta_j) = 0, \qquad \frac{1}{M_j}\sum_{k=1}^{M_j} q_{jk}(\theta_j)^2 = 1, \tag{13}$$

which fixes location and scale and ensures ordinal scores are comparable to the standardized continuous variables.

We consider three flexible two-parameter families for $Q(\cdot;\theta_j)$, all obtained by a monotone transformation of a standard normal Z:

Johnson S_U distribution [26]:

$$Y = \sinh\left(\frac{Z-\gamma}{\delta}\right), \quad \theta_j = (\gamma,\delta),\ \delta > 0, \tag{14}$$

where γ governs skewness and δ tail heaviness; $\gamma = 0$ gives a distribution symmetric about 0.

Tukey g-and-h distribution [27]:

$$Y = \begin{cases} Z\exp(hZ^2/2), & g = 0, \\ \dfrac{e^{gZ}-1}{g}\exp(hZ^2/2), & g \neq 0, \end{cases} \quad \theta_j = (g,h),\ h \geq 0, \tag{15}$$

where g controls skewness and h tail heaviness independently; $g = h = 0$ reduces to the standard normal.

Sinh-arcsinh distribution [28]:

$$Y = \sinh\left(\frac{\operatorname{arcsinh}(Z)+\varepsilon}{\delta}\right), \quad \theta_j = (\varepsilon,\delta),\ \delta > 0, \tag{16}$$

where ε governs skewness and δ tail weight; $\delta > 1$ yields lighter tails than the normal, $\delta < 1$ heavier; $\varepsilon = 0$, $\delta = 1$ gives the identity.

All three families represent a wide range of skewness and kurtosis combinations, making them suitable surrogates for the unknown true spacing of ordinal levels.

3.3.2. Data-Driven Estimation of θ_j

In ref. [13], θ_j is chosen by minimizing the residual deviance of a regression model. In an unsupervised context, no response variable is available. We instead propose a *two-stage* strategy that aligns the estimation criterion with the clustering objective.

Stage 1 — Provisional partition. Run a fast preliminary k-means (five iterations, equal weights, integer ordinal codes, $\tau = \frac{1}{2}$) to obtain an initial partition $C^{(0)} = (C_1^{(0)}, \ldots, C_K^{(0)})$.
Stage 2 — Score estimation. For each ordinal variable j independently, estimate θ_j by maximizing the variable-wise adjusted Davies–Bouldin criterion at the provisional partition:

$$\hat{\theta}_j = \arg\max_{\theta_j \in \Theta} \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k} \frac{\theta_{kk',j}(C^{(0)},\theta_j)}{\sigma_{k,j}(C^{(0)},\theta_j) + \sigma_{k',j}(C^{(0)},\theta_j)}, \tag{17}$$

where $\sigma_{k,j}$ and $\theta_{kk',j}$ are evaluated using $d_{ii',j} = (q_{j,x_{ij}}(\theta_j) - q_{j,x_{i'j}}(\theta_j))^2$ (plain squared score difference at initialization, $\tau = \frac{1}{2}$). Because $\Theta \subset \mathbb{R}^2$ is low-dimensional and the criterion is C^∞ in θ_j, this is solved by Nelder–Mead at negligible computational cost relative to the main loop.

Once $\hat{\theta}_j$ is obtained for every ordinal variable, the scores $q_{jk}(\hat{\theta}_j)$ are fixed and the dissimilarities $d^{(O)}_{ii',j}(\tau,\hat{\theta}_j)$ of (10) enter the main optimization only through the pre-estimated

scores, with τ remaining a free parameter. This two-stage design decouples score estimation from sparse weight optimization, preserving the theoretical guarantees of Section 4: since the scores are fixed constants in the main loop, the concentration structure of $d^{(O)}_{ii',j}(\tau, \hat{\theta}_j)$ is identical to that of the continuous dissimilarity, with the factor $c(\tau, \cdot)$ absorbed into the bounding constants of Lemma A1. The asymptotic validity of this plug-in step is discussed in Appendix A.

3.4. Dissimilarity for Nominal Variables

Nominal variables are unordered categorical variables for which no metric structure exists. We adopt the co-occurrence dissimilarity of [29] (Equations (18)–(20)), which measures the distance between two levels of a nominal variable by comparing, across all other nominal variables, the conditional distributions induced by each level. For a nominal variable j with levels $\{1, \dots, L_j\}$, the raw distance between levels a and b is

$$\delta_{ab}(j) = \sum_{j' \neq j} \sum_{l'=1}^{L_{j'}} \max\{\hat{\mathbb{P}}(X_{j'} = l' \mid X_j = a),\ \hat{\mathbb{P}}(X_{j'} = l' \mid X_j = b)\}, \tag{18}$$

where $\hat{\mathbb{P}}(X_{j'} = l' \mid X_j = a) = \#\{i : x_{ij} = a,\, x_{ij'} = l'\}/\#\{i : x_{ij} = a\}$. Here, q is the number of nominal variables. Each of the $q - 1$ other variables contributes a term in $[1, 2]$, so the raw distance ranges in $[q - 1, 2(q - 1)]$ and is normalized to $[0, 1]$:

$$\tilde{\delta}_{ab}(j) = \frac{\delta_{ab}(j)}{q - 1} - 1. \tag{19}$$

The pairwise dissimilarity between observations i and i' on nominal variable j is then

$$d^{(N)}_{ii',j} = \sum_{l=1}^{L_j} \sum_{\substack{l'=1 \\ l' \neq l}}^{L_j} \tilde{\delta}_{ll'}(j)\ \mathbf{1}(x_{ij} = l)\ \mathbf{1}(x_{i'j} = l'). \tag{20}$$

This formulation assigns small distances to levels that induce similar conditional distributions on all other nominal variables, thereby capturing inter-variable dependence structure that simple matching coefficients ignore. The dissimilarity lies in $[0, 1]$ almost surely, a property used in the Bernstein concentration argument of Appendix A.

3.5. Unified Weighted Dissimilarity

The three type-specific dissimilarities are combined into a global weighted dissimilarity. Because continuous and ordinal variables both contribute $c(\tau, \cdot) \cdot (\cdot)^2$-type terms on a comparable normalized scale, they share a single weight vector $\mathbf{w}$. Nominal variables, whose co-occurrence dissimilarity has a different range and structure, are assigned a separate weight vector $\mathbf{z}$. The full unified dissimilarity is

$$d(i, i'; \mathbf{w}, \mathbf{z}, \tau, \hat{\boldsymbol{\theta}}) = \underbrace{\sum_{j=1}^{p} w_j\, d^{(C)}_{ii',j}(\tau) + \sum_{j=p+1}^{p+r} w_j\, d^{(O)}_{ii',j}(\tau, \hat{\theta}_j)}_{\text{metric variables (continuous + ordinal)}} + \underbrace{\sum_{j=p+r+1}^{p+r+q} z_j\, d^{(N)}_{ii',j}}_{\text{nominal variables}}, \tag{21}$$

where $\mathbf{w} = (w_1, \dots, w_{p+r})^\top \geq \mathbf{0}$ and $\mathbf{z} = (z_{p+r+1}, \dots, z_{p+r+q})^\top \geq \mathbf{0}$.

Equation (21) has a clear two-group interpretation. All metric variables share the single global asymmetry coefficient τ: continuous variables through their raw differences and ordinal variables through their Azzalini score differences. This is the natural consequence of the fact that, after normalization, both variable types lie on the same real-valued scale and contribute the same squared-type loss. Separating the two weight vectors $\mathbf{w}$ and $\mathbf{z}$ further prevents the co-occurrence scale of nominal dissimilarities from distorting the weighting of the metric variables.

The weights are not specified a priori but are learned simultaneously with the partition and with τ, as detailed in Section 4.1. This data-driven weighting automatically sets to zero the weights of uninformative variables while amplifying the contribution of variables that genuinely discriminate between clusters.

3.6. Cluster Dispersion Criteria

For a partition $C = (C_1, \dots, C_K)$ with $|C_k| = n_k$, and for a generic pairwise dissimilarity $d_{ii',j}$ (which may be any of the three type-specific dissimilarities above), the within-cluster dispersion of variable j in cluster k is

$$\sigma_{k,j}(C) = \frac{1}{n_k(n_k-1)} \sum_{\substack{i,i' \in C_k \\ i \neq i'}} d_{ii',j}, \tag{22}$$

and the between-cluster dispersion between clusters k and k' on variable j is

$$\theta_{kk',j}(C) = \frac{1}{n_k\, n_{k'}} \sum_{i \in C_k} \sum_{i' \in C_{k'}} d_{ii',j}. \tag{23}$$

These quantities measure, respectively, the compactness of each cluster and the separation between cluster pairs with respect to a single variable. They form the building blocks of the variable-wise discriminative score used in the sparse optimization of Section 4.1. The within- and between-cluster dispersions follow [15].

Remark 1 (Ordered pairs and asymmetry)**.** *For a metric variable j with $d_{ii',j} = d_\tau(x_{ij} - x_{i'j})$ (continuous) or d_τ applied to scores (ordinal), identity (8) gives*

$$\sigma_{k,j}(C) = \frac{2}{n_k(n_k-1)} \sum_{\substack{i,i' \in C_k \\ i<i'}} (x_{ij} - x_{i'j})^2 = \frac{2n_k}{n_k - 1}\, v_{k,j}, \tag{24}$$

where $v_{k,j}$ is the within-cluster variance. The within-cluster dispersion (22), which sums over ordered pairs, is therefore symmetric and independent of τ. In contrast, the between-cluster dispersion (23) sums over $i \in C_k$ and $i' \in C_{k'}$ in one fixed order, so, in general, $\theta_{kk',j} \neq \theta_{k'k,j}$, and τ acts only through $\theta_{kk',j}$. In adjDBI_j*, the pair (k, k') is always evaluated with $i \in C_k$ and $i' \in C_{k'}$.*

4. Sparse Clustering and Theoretical Properties

4.1. General Sparse Clustering Framework

The sparse clustering paradigm seeks to jointly estimate a partition of the n observations and identify the subset of variables that genuinely contribute to the cluster structure. Following [14], the general formulation assigns a non-negative weight w_j to each variable j and maximizes a weighted sum of variable-wise cluster-quality scores, subject to constraints that drive irrelevant weights to zero:

$$\max_{\mathbf{w},\, \boldsymbol{\Theta} \in \mathcal{D}} \sum_{j=1}^{p+r+q} w_j f(X_j, \boldsymbol{\Theta}), \qquad \text{subject to} \quad \|\mathbf{w}\|_2^2 \leq 1, \quad \|\mathbf{w}\|_1 \leq s, \quad w_j \geq 0, \tag{25}$$

where $f(X_j, \boldsymbol{\Theta})$ measures the contribution of variable j to the clustering $\boldsymbol{\Theta}$, and $s \in [1, \sqrt{p+r+q}]$ controls sparsity. The ℓ_1 constraint is the key ingredient: when s is small, the optimal $\mathbf{w}$ concentrates on few variables, setting irrelevant weights to zero.

In the mixed-type setting of this paper, three variable types require three distinct contribution functions. Following [15], we therefore use two separate weight vectors: $\mathbf{w} = (w_1, \dots, w_{p+r})^\top$ for metric variables (continuous and Azzalini-scored ordinal), and $\mathbf{z} = (z_{p+r+1}, \dots, z_{p+r+q})^\top$ for nominal variables, with independent sparsity bounds s_w and s_z. This separation prevents the co-occurrence scale of nominal dissimilarities from distorting the weighting of the metric squared-type dissimilarities.

4.2. Objective Function

The unified pairwise dissimilarity derived in Section 3.5 is

$$d(X_i, X_{i'}; \mathbf{w}, \mathbf{z}, \tau, \hat{\boldsymbol{\theta}}) = \sum_{j=1}^{p} w_j \, d^{(C)}_{ii',j}(\tau) + \sum_{j=p+1}^{p+r} w_j \, d^{(O)}_{ii',j}(\tau, \hat{\theta}_j) + \sum_{j=p+r+1}^{p+r+q} z_j \, d^{(N)}_{ii',j}, \tag{26}$$

where $d^{(C)}_{ii',j}(\tau)$, $d^{(O)}_{ii',j}(\tau, \hat{\theta}_j)$, and $d^{(N)}_{ii',j}$ are the expectile-based continuous, expectile-based ordinal, and co-occurrence dissimilarities defined in Sections 3.2–3.4, respectively. The expectile parameter τ appears in both metric terms: it acts on raw differences for continuous variables and on Azzalini score differences for ordinal variables. The ordinal scores $q_{jk}(\hat{\theta}_j)$ are fixed from the two-stage estimation of Section 3.3.2 and treated as known constants throughout.

Clustering quality at the variable level is assessed through the adjusted Davies–Bouldin index [15,30]. For a partition $C = (C_1, \ldots, C_K)$ and a generic variable-j dissimilarity $d_{ii',j}$, define

$$\text{adjDBI}_j(C) = \frac{1}{K} \sum_{k=1}^{K} \max_{k' \neq k} \frac{\theta_{kk',j}(C)}{\sigma_{k,j}(C) + \sigma_{k',j}(C)}, \tag{27}$$

with $\sigma_{k,j}$ and $\theta_{kk',j}$ as defined in (22) and (23). A larger value of adjDBI_j signals that variable j separates at least one cluster pair more sharply than the clusters are spread internally. This reverses the original DBI convention (where lower is better) so that maximization over weights is natural.

Substituting (27) as the contribution function in (25), and using superscripts (w) and (z) to distinguish the dissimilarity used for the two variable groups, we obtain the full ExSC-$\tau\theta$ objective:

$$\max_{\substack{C_1,\ldots,C_K \\ \mathbf{w},\mathbf{z},\tau}} \left\{ \sum_{j=1}^{p+r} w_j \, \text{adjDBI}^{(w)}_j(C, \tau, \hat{\boldsymbol{\theta}}) \; + \sum_{j=p+r+1}^{p+r+q} z_j \, \text{adjDBI}^{(z)}_j(C) \right\} \tag{28}$$

subject to

$$\|\mathbf{w}\|_2^2 \leq 1, \quad \|\mathbf{z}\|_2^2 \leq 1, \quad \|\mathbf{w}\|_1 \leq s_w, \quad \|\mathbf{z}\|_1 \leq s_z, \quad w_j, z_j \geq 0 \; \forall j, \quad \tau \in (0,1). \tag{29}$$

The sparsity parameters satisfy $1 \leq s_w \leq \sqrt{p+r}$ and $1 \leq s_z \leq \sqrt{q}$. Setting $s_w = \sqrt{p+r}$ (resp. $s_z = \sqrt{q}$) places all weight on the unit sphere without enforcing sparsity; smaller values increasingly concentrate weight on fewer variables. All parameters of the algorithm, with their role, admissible values and selection rule, are summarized in Table 3.

Table 3. Parameters of ExSC-$\tau\theta$: Role, admissible values and selection.

Symbol	Role	Domain	Selected by
K	number of clusters	integers ≥ 2	minimum mDBI (46)
s_w	ℓ_1 bound on metric weights	$[1, \sqrt{p+r}]$	minimum mDBI
s_z	ℓ_1 bound on nominal weights	$[1, \sqrt{q}]$	minimum mDBI
τ	asymmetry of d_τ	$[\tau_{\min}, \tau_{\max}] \subset (0,1)$	golden-section search
θ_j	shape of the ordinal scores of variable j	$\Theta \subset \mathbb{R}^2$ (family-specific)	Nelder–Mead on (17)
λ	soft-threshold level in Steps 2–3	≥ 0	from $\|\mathbf{w}\|_1 = s_w$
tol, T_0	stopping tolerance; preliminary k-means iterations	10^{-4}; 5	fixed

Problem (28) and (29) is solved by the coordinate descent of Algorithm 1, which alternates between updating C (weighted k-means), $\mathbf{w}$ and $\mathbf{z}$ (soft-thresholding), and τ (golden-section search). The tuning parameters K, s_w, and s_z are chosen by minimizing the modified Davies–Bouldin index of (46) over a candidate grid.

Variable Selection in ExSC-$\tau\theta$

Variable selection is *embedded*: it is performed by the ℓ_1/ℓ_2-constrained weights $\mathbf{w}, \mathbf{z}$ of (28) and (29), in the sense of [14], and not by a filter, wrapper or meta-heuristic search. The weight of a variable depends on its univariate score adjDBI_j; consequently, interactions between variables are not modeled, the sparsity bounds s_w, s_z must be tuned, and metric and nominal variables are selected in two separate groups. Filter, wrapper or meta-heuristic selectors would act independently of the clustering objective and are outside the scope of the theory of this paper. Principal component analysis is not used for dimension reduction or selection: it is used in Section 6 only to visualize the final partitions in two dimensions, and all clustering is carried out in the original variable space.

4.3. Population-Level Analysis and Identifiability

To establish consistency, we work at the population level and introduce the objects governing the large-sample behavior of the estimator.

4.3.1. Population Discriminative Scores

For a partition C and variable j, the population discriminative score is

$$\overline{\alpha}_j(C) = \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k} \frac{\mathbb{E}[\theta_{kk',j}(C)]}{\mathbb{E}[\sigma_{k,j}(C)] + \mathbb{E}[\sigma_{k',j}(C)]}, \tag{30}$$

and the population objective is

$$E(C, \mathbf{w}, \mathbf{z}) = \sum_{j=1}^{p+r} w_j \, \overline{\alpha}_j^{(w)}(C) + \sum_{j=p+r+1}^{p+r+q} z_j \, \overline{\alpha}_j^{(z)}(C). \tag{31}$$

The empirical counterpart (28) is denoted $F(C, \mathbf{w}, \mathbf{z})$ throughout.

4.3.2. Signal–Noise Separation

Let C^* denote the true partition. A variable j is a signal variable if $\overline{\alpha}_j(C^*)$ is bounded away from the noise level, and a noise variable otherwise. Formally, there exists $\Delta > 0$ such that

$$\overline{\alpha}_j(C^*) > \Delta \ \text{ for all signal variables } j, \qquad \overline{\alpha}_j(C^*) \leq \Delta \ \text{ for all noise variables } j. \tag{32}$$

This is the unsupervised analogue of the beta-min condition in high-dimensional regression [31]. It ensures that the population-optimal weights place all mass on signal variables, making the true support identifiable.

4.3.3. Population-Optimal Weights

Under (32), the population-optimal weight for the j-th metric signal variable is

$$w_j^* = \frac{\overline{\alpha}_j^{(w)}(C^*) - \Delta}{\Big(\sum_{j'=1}^{(p+r)^*} \big(\overline{\alpha}_{j'}^{(w)}(C^*) - \Delta\big)^2\Big)^{1/2}}, \tag{33}$$

where $(p+r)^*$ is the number of signal metric variables; z_j^* is defined analogously for nominal signals. These weights are strictly positive for signal variables and exactly zero for noise variables.

The following two lemmas underpin the consistency proof; their proofs are given in Appendix A.3.

Lemma 1 (Hölder control of the objective)**.** *Let* $\alpha_j(C) = \mathrm{adjDBI}_j^{(w)}(C, \tau, \hat{\boldsymbol{\theta}})$, $F(C, \mathbf{w}) = \sum_{j=1}^{p+r} w_j \alpha_j(C)$ *and* $E(C, \mathbf{w}) = \sum_{j=1}^{p+r} w_j \bar{\alpha}_j(C)$*. Then, for any partition* C *and any two feasible weight vectors* $\mathbf{w}, \mathbf{w}'$*,*

$$\begin{aligned} |F(C, \mathbf{w}) - F(C, \mathbf{w}')| &\le \|\alpha(C)\|_\infty \, \|\mathbf{w} - \mathbf{w}'\|_1, \\ |F(C, \mathbf{w}) - E(C, \mathbf{w})| &\le \|\mathbf{w}\|_1 \max_j |\alpha_j(C) - \bar{\alpha}_j(C)| \\ &\le s_w \max_j |\alpha_j(C) - \bar{\alpha}_j(C)|. \end{aligned} \tag{34}$$

Remark 2. *No uniform bound on* $\|\alpha(C)\|_\infty$ *is claimed:* adjDBI_j *is a between/within ratio and increases with the separation of the cluster means. The proof of Theorem 1 uses only the second inequality of Lemma 1, i.e., (A15).*

Lemma 2 (Identifiability of the population maximizer)**.** *Under condition (32) and the constraints (29) with* s_w *in the range of Theorem 1, the pair* $(C^*, \mathbf{w}^*)$ *is the unique maximizer of* $E(C, \mathbf{w}, \mathbf{z}^*)$ *over the feasible set. Moreover, there exists* $\delta > 0$ *such that for every feasible* $(C, \mathbf{w}) \neq (C^*, \mathbf{w}^*)$*,*

$$E(C^*, \mathbf{w}^*) - E(C, \mathbf{w}) \ge \delta. \tag{35}$$

4.4. Data-Generating Model

We specify a mixture model governing the theoretical analysis. Let $\phi_i \in \{1, \dots, K\}$ denote the latent cluster label of observation i, drawn from a multinomial with probabilities $(\pi_1, \dots, \pi_K)$, $\pi_k > 0$ for all k.

4.4.1. Continuous Variables

Conditional on $\phi_i = k$, $X_i^{(1)} \mid \phi_i = k \sim \mathcal{N}(\mu_k, I_p)$, where $\mu_k = (\mu_{k1}, \dots, \mu_{kp})^\top$. Variable $j \le p^*$ is a signal variable if $\mu_{kj} \neq \mu_{k'j}$ for some $k \neq k'$; noise variables satisfy $\mu_{kj} = 0$ for all k.

4.4.2. Ordinal Variables

Conditional on $\phi_i = k$, the j-th ordinal variable follows a multinomial over $\{1, \dots, M_j\}$ with probabilities $\mathbf{p}_{kj}^{(O)}$. Variable j is a signal ordinal variable ($j - p \le r^*$) if there exist $k \neq k'$ such that

$$\sum_{\ell=1}^{M_j} \min\{p_{kj}^{(O)}(\ell),\ p_{k'j}^{(O)}(\ell)\} \le 0.4, \tag{36}$$

requiring a total-variation distance of at least 0.6 between cluster level distributions. Noise ordinal variables are uniform over their levels. After the pre-estimation described in Section 3.3.2, the scores $q_{jk}(\hat{\theta}_j)$ are treated as deterministic constants since $\hat{\theta}_j \xrightarrow{p} \theta_j^*$ under (36).

4.4.3. Nominal Variables

Conditional on $\phi_i = k$, the j-th nominal variable is multinomial over $\{1, \dots, L_j\}$ with probabilities $\mathbf{p}_{kj}^{(N)}$. Variable j is a signal nominal variable ($j - p - r \le q^*$) if (36) holds with $\mathbf{p}_{kj}^{(N)}$ in place of $\mathbf{p}_{kj}^{(O)}$. Noise nominal variables are uniform.

4.4.4. Discriminative Bounds

Define the variable-level signal strengths

$$\beta_{kj}^{(1)} = \max_{k' \neq k} (\mu_{kj} - \mu_{k'j})^2, \tag{37}$$

$$\beta_{kj}^{(2)} = \max_{k' \neq k} \frac{\sum_{\ell,\ell'} (p_{kj}^{(N)}(\ell) - p_{k'j}^{(N)}(\ell))(p_{k'j}^{(N)}(\ell') - p_{kj}^{(N)}(\ell'))\, \tilde{\delta}_{\ell\ell'}(j)}{\sum_{\ell,\ell'} p_{kj}^{(N)}(\ell) p_{kj}^{(N)}(\ell') \tilde{\delta}_{\ell\ell'}(j) + \sum_{\ell,\ell'} p_{k'j}^{(N)}(\ell) p_{k'j}^{(N)}(\ell') \tilde{\delta}_{\ell\ell'}(j)}, \tag{38}$$

and let

$$\xi_1 = \min_{k,j} \beta_{kj}^{(1)}, \quad \xi_2 = \max_{k,j} \beta_{kj}^{(1)}, \quad \xi_3 = \min_{k,j} \beta_{kj}^{(2)}, \quad \xi_4 = \max_{k,j} \beta_{kj}^{(2)}. \tag{39}$$

Signal variables satisfy $\xi_1, \xi_3 > 0$ by construction. The expectile parameter τ enters the population analysis through the factor $c(\tau, \cdot) \in \{2\tau, 2(1-\tau)\}$ in the dissimilarity; its effect is bounded uniformly over $\tau \in (0,1)$ and absorbed into the constant $(1+\xi_2)$ of Lemma A1.

4.5. Screening Consistency

Theorem 1 (Screening consistency of ExSC-$\tau\theta$)**.** *Let $(\hat{C}, \hat{\mathbf{w}}, \hat{\mathbf{z}})$ be the solution to (28) and (29) obtained by Algorithm 1, with ordinal scores pre-estimated as in Section 3.3.2. Define the population mean discriminative scores $\overline{\alpha}_{w,j} = (2K)^{-1} \sum_k \beta_{kj}^{(1)}$ for $j \leq p + r$ and $\overline{\alpha}_{z,j} = (2K)^{-1} \sum_k \beta_{kj}^{(2)}$ for $j > p + r$. Suppose the sparsity parameters satisfy*

$$\frac{\sum_{j=1}^{(p+r)^*} \overline{\alpha}_{w,j} - \frac{1}{4}\xi_1 (p+r)^*}{\left(\sum_{j=1}^{(p+r)^*} \left(\overline{\alpha}_{w,j} - \frac{1}{4}\xi_1 \right)^2 \right)^{1/2}} \leq s_w \leq \frac{\sum_{j=1}^{(p+r)^*} \overline{\alpha}_{w,j}}{\left(\sum_{j=1}^{(p+r)^*} \overline{\alpha}_{w,j}^2 \right)^{1/2}}, \tag{40}$$

$$\frac{\sum_{j=1}^{q^*} \overline{\alpha}_{z,j} - \frac{1}{4}\xi_3 q^*}{\left(\sum_{j=1}^{q^*} \left(\overline{\alpha}_{z,j} - \frac{1}{4}\xi_3 \right)^2 \right)^{1/2}} \leq s_z \leq \frac{\sum_{j=1}^{q^*} \overline{\alpha}_{z,j}}{\left(\sum_{j=1}^{q^*} \overline{\alpha}_{z,j}^2 \right)^{1/2}}. \tag{41}$$

Assume furthermore that

$$\left((p+r)^*\right)^2 \leq \frac{\xi_1^4}{2048\, \xi_2^2\, (1+\xi_2) \ln K}, \tag{42}$$

$$(q^*)^2 \leq \frac{\xi_3^4}{2048\, \xi_4^2\, (1+\xi_4) \ln K}, \tag{43}$$

$$\ln(p+r+q) = o(n). \tag{44}$$

Then, as $n \to \infty$,

$$\mathbb{P}(\{1, \ldots, (p+r)^*\} \subseteq \mathrm{supp}(\hat{\mathbf{w}})) \to 1, \qquad \mathbb{P}(\{p+r+1, \ldots, (p+r+q)^*\} \subseteq \mathrm{supp}(\hat{\mathbf{z}})) \to 1. \tag{45}$$

The full proof is given in Appendix A.

Remark 3 (Dimensionality conditions)**.** *Conditions (42) and (43) bound the number of signal variables in terms of the signal-to-noise ratios ξ_1/ξ_2 and ξ_3/ξ_4, analogously to the sparsity conditions in high-dimensional regression [31]. They are mild: when $\xi_1 = \xi_2$, condition (42) reduces to $(p+r)^* \leq c/\sqrt{\ln K}$ for a constant c depending only on ξ_1. Condition (44), $\ln(p+r+q) = o(n)$, permits the total dimension to grow subexponentially in n, accommodating the high-dimensional settings targeted by the method. The roles of τ and of the plug-in $\hat{\boldsymbol{\theta}}$ in the proof are detailed in Appendix A.*

Remark 4 (Scope of Theorem 1)**.** *Theorem 1 is a* variable-screening *result:* $\mathbb{P}(S^* \subseteq \mathrm{supp}(\hat{\mathbf{w}})) \to 1$. *It does not state that* $\hat{C} \to C^*$. *Hence, the primary task of the method, clustering accuracy, has no large-sample guarantee in this paper, only a variable-selection guarantee; clustering performance is assessed empirically in Sections 5 and 6. Partition-recovery rates for the expectile criterion are left for future work.*

4.6. Algorithm and Computational Complexity

Step 1: Weighted k-means. With $(\mathbf{w}, \mathbf{z}, \tau)$ fixed, the partition is updated by weighted k-means in the sense of sparse k-means [14] (package `sparcl`, 100 random starts): Lloyd's iterations on a numerical representation in which each metric variable j (continuous, or ordinal after scoring) is multiplied by $\sqrt{w_j}$, and each nominal variable j is expanded by dummy (one-hot) coding into L_j columns multiplied by $\sqrt{z_j}$. Lloyd's iterations require a symmetric, centroid-based criterion, which is available here: by identity (8), the symmetric part $\frac{1}{2}[d(i,i') + d(i',i)]$ of the metric dissimilarity equals $\sum_j w_j (x_{ij} - x_{i'j})^2$, i.e., the weighted squared Euclidean distance, with cluster means as centroids. The asymmetric dissimilarity d_τ is therefore *not* used in the assignment step. The co-occurrence dissimilarity (20) is used in the variable-wise scores $\mathrm{adjDBI}_j^{(z)}$ of nominal variables (Step 3 and criterion (46)); the dummy coding is used only in this step. The asymmetry parameter τ acts through $\theta_{kk',j}$ in adjDBI_j and hence through the weights $\mathbf{w}$ (Steps 2 and 4), which in turn reshape the weighted k-means. The preliminary Stage-1 partition (Section 3.3.2) is the same k-means with equal weights, integer ordinal codes and $\tau = \frac{1}{2}$ (5 iterations).

Tuning parameter selection. The cluster number K and the sparsity bounds (s_w, s_z) are chosen by minimizing a modified Davies–Bouldin index over a candidate grid:

$$\mathrm{mDBI} = \frac{1}{K} \sum_{k=1}^{K} \left(\frac{\sum_{j=1}^{p+r} w_j \min_{k' \neq k} \frac{\sigma_{k,j} + \sigma_{k',j}}{\theta_{kk',j}}}{\sum_{j=1}^{p+r} w_j} + \frac{\sum_{j=p+r+1}^{p+r+q} z_j \min_{k' \neq k} \frac{\sigma_{k,j} + \sigma_{k',j}}{\theta_{kk',j}}}{\sum_{j=p+r+1}^{p+r+q} z_j} \right). \tag{46}$$

Note that mDBI uses the standard DBI ratio $(\sigma_{k,j} + \sigma_{k',j})/\theta_{kk',j}$ (smaller values indicate better clustering), which is the inverse of the adjDBI_j criterion used in the objective (28); minimizing mDBI over the grid is therefore consistent with maximizing (28).

Computational complexity. Let $\bar{q} = \sum_j L_j$ be the number of dummy columns of the nominal variables. Per iteration, (i) the k-means update costs $O(T_{km}\, nK(p + r + \bar{q}))$, with T_{km} Lloyd iterations; (ii) the within-cluster dispersions of metric variables have the closed form of Remark 1, $O(n)$ per variable, whereas the between-cluster dispersions $\theta_{kk',j}$ involve every cross-cluster pair once, $O(n^2)$ per variable (also for nominal variables); (iii) Steps 2–3 are $O(p + r + q)$ given the scores; and (iv) Step 4 (golden-section search over τ) requires $O(\log \varepsilon^{-1})$ evaluations of the metric $\theta_{kk',j}$, each $O(n^2(p + r))$. The co-occurrence dissimilarities and the score estimation (Nelder–Mead calls in $\mathbb{R}^2$, $O(n)$ per criterion evaluation) are computed once before the main loop and are negligible. Overall, the method scales as $O(Tn^2(p + r + q) + T\, T_{km}\, nK(p + r + \bar{q}))$, i.e., $O(Tn^2(p + r + q))$ when n dominates $K\, T_{km}$, where T is the number of iterations until convergence. Table 4 compares the per-iteration time complexity of ExSC-$\tau\theta$ with that of the other methods applicable to mixed-type data. ExSC-$\tau\theta$ is quadratic in n, like DBI-SC, and more expensive than methods that are linear in n; its observed advantage over DBI-SC (Table 15) is a constant-factor effect.

Table 4. Time complexity per iteration of methods applicable to mixed-type data (T: number of iterations; M: number of variable subsets visited by VarSelLCM).

Method	Cost per Iteration	Variable Selection
Sparse k-means [14] (continuous only)	$O(nKp)$	ℓ_1 weights
VarSelLCM [4]	$O(nK(p+r+q))$ per EM step, times M	MICL
DBI-SC [15]	$O(n^2(p+r+q))$	ℓ_1 weights
ExSC-$\tau\theta$	$O(n^2(p+r+q))$	ℓ_1 weights

4.7. Illustrative Example

To make the mechanics of ExSC-$\tau\theta$ concrete, we follow the whole procedure on a deliberately small dataset. The example is an illustration of the steps of Section 3 and Algorithm 1, not statistical evidence of performance, which is the object of Sections 5 and 6.

4.7.1. Data

A retailer describes $n = 12$ customers by five variables (Table 5): two continuous variables, annual income (in thousands of dirhams) and age; one ordinal variable, satisfaction, with the ordered levels 1 = Low, 2 = Medium, 3 = High, and 4 = Very high; and two nominal variables, region (North, Center, South) and purchase channel (web, shop). The data were built so that two groups of six customers differ in income, satisfaction and region, while age and channel are unrelated to the groups. The task is to recover the two groups *and* to discover which variables matter. We fix $K = 2$, $s_w = 1.2$ and $s_z = 1$ (so that one nominal variable is retained), and restrict $\theta = (\varepsilon, \delta)$ of the sinh-arcsinh family to $[-3, 3] \times [0.4, 3]$.

Table 5. Toy mixed-type dataset ($n = 12$). Income is in thousands of dirhams. The last column gives the cluster returned by ExSC-$\tau\theta$.

ID	Income	Age	Satisfaction	Region	Channel	Cluster
1	24	44	1 (Low)	North	web	1
2	31	29	2 (Medium)	North	shop	1
3	28	53	1 (Low)	North	shop	1
4	36	35	2 (Medium)	North	web	1
5	27	61	3 (High)	Center	web	1
6	41	38	2 (Medium)	North	shop	1
7	52	49	4 (Very high)	South	shop	2
8	47	31	3 (High)	South	web	2
9	63	57	4 (Very high)	South	web	2
10	58	26	4 (Very high)	Center	shop	2
11	71	42	4 (Very high)	South	shop	2
12	45	58	3 (High)	South	web	2

4.7.2. Step 1: Preprocessing

Income and age are standardized to zero mean and unit variance (Section 3.2); for instance, customer 1 has standardized income -1.345 and customer 7 has 0.578. Satisfaction is kept as an ordered factor with four levels, and region and channel as integer-coded factors.

4.7.3. Step 2: Learned Scores for the Ordinal Variable

A preliminary k-means with equal weights and integer codes gives the provisional partition $\{1, \dots, 6\}$ and $\{7, \dots, 12\}$. The parameters θ of the sinh-arcsinh family are then chosen by maximizing the variable-wise adjusted Davies–Bouldin criterion (17) at this partition, which gives the normalized scores of Table 6. The learned scores are *not* equally spaced: the step from High to Very high (1.19) is much larger than the step from Low to Medium (0.64), whereas integer coding imposes the same step (0.90) between consecutive levels. The criterion increases from 2.43 with standardized integer codes to 2.58 with the learned scores. (The maximizer is found on the boundary $\varepsilon = 3$ of the search box where

the criterion is almost flat; the resulting scores are stable to two decimals when the box is enlarged, so the scores, rather than $\hat{\theta}$ itself, are the quantity of interest.)

Table 6. Numerical scores for the four satisfaction levels (both coding schemes satisfy normalization (9)).

	Low	Medium	High	Very High
Integer codes, standardized	−1.342	−0.447	0.447	1.342
Learned scores $q_{jk}(\hat{\theta}_j)$	−1.198	−0.560	0.286	1.473

4.7.4. Step 3: The Three Dissimilarities

For the nominal variables, the co-occurrence dissimilarity (18)–(20) is computed from the data: with only two nominal variables, the distance between two regions is obtained by comparing their distributions of channel, which gives $\tilde{\delta}(\text{North}, \text{Center}) = 0.10$, $\tilde{\delta}(\text{Center}, \text{South}) = 0.10$ and $\tilde{\delta}(\text{North}, \text{South}) = 0.20$; the distance between the channels web and shop, obtained from the distribution of region, is 0.17. The algorithm returns $\hat{\tau} = 0.613$, so positive gaps on the metric variables are weighted by $2\hat{\tau} = 1.227$ and non-positive gaps by $2(1 - \hat{\tau}) = 0.773$. Table 7 gives the dissimilarities for three ordered pairs: customers 1 and 2 (same group), customers 1 and 7 (different groups), and the same pair in the reverse order. For the pair $(1, 7)$, the standardized income gap is -1.92, so $d_{\hat{\tau}} = 0.773 \times 1.92^2 = 2.86$, whereas for $(7, 1)$, the gap is $+1.92$ and $d_{\hat{\tau}} = 1.227 \times 1.92^2 = 4.53$; the two values sum to $2u^2 = 7.39$, as stated by identity (8). The unified weighted dissimilarity (21) uses the weights of Step 4.

The table shows how the pieces fit together: income and satisfaction dissimilarities are small within a group and large across groups, region adds a small contribution, and age and channel show no increase across groups (the age dissimilarity is even larger for the same-group pair), so they carry no group information. The asymmetry appears only in the metric columns, and only when the order of the two customers is reversed.

Table 7. Type-specific dissimilarities for three ordered pairs of customers, and the unified weighted dissimilarity (21) with the weights of Table 8.

Pair (i, i')	Income	Age	Satisfaction	Region	Channel	$d(i, i')$
$(1, 2)$ same group	0.179	2.074	0.315	0.000	0.167	0.245
$(1, 7)$ different groups	2.859	0.145	5.518	0.200	0.167	4.231
$(7, 1)$ reversed order	4.535	0.230	8.753	0.200	0.167	6.594

Table 8. Variable-wise scores adjDBI_j and sparse weights returned by ExSC-$\tau\theta$ on the toy data.

Variable	Type	adjDBI_j	Weight
Income	continuous	2.64	$w = 0.974$
Age	continuous	0.42	$w = 0$
Satisfaction	ordinal	2.58	$w = 0.226$
Region	nominal	2.50	$z = 1$
Channel	nominal	0.42	$z = 0$

4.7.5. Steps 4–5: Variable Scores, Sparse Weights and Partition

Algorithm 1 alternated the weighted k-means update, the soft-thresholded weights and the update of τ, and converged in two iterations. Table 8 gives the variable-wise adjusted Davies–Bouldin scores and the resulting weights. Income, satisfaction and region obtain large scores (2.64, 2.58 and 2.50); age and channel obtain small ones (0.42 each). Under the ℓ_1 bound $s_w = 1.2$, the metric weights are $(0.974, 0, 0.226)$ for (income, age, satisfaction), so age is set *exactly* to zero, and under $s_z = 1$, the nominal weights are $(1, 0)$, so channel is also

set to zero. The final partition is $\{1, \ldots, 6\}$ and $\{7, \ldots, 12\}$ (Table 5), which coincides with the groups used to build the data (adjusted Rand index equal to 1); the mean incomes of the two clusters are 31.2 and 56.0 thousand dirhams.

On such a small and clean example, an unweighted k-means with integer codes also recovers the two groups, so the point of the illustration is not the partition alone: ExSC-$\tau\theta$ additionally tells the analyst which variables define the groups (income, satisfaction, region), sets the others exactly to zero, assigns the ordinal levels a data-driven spacing in which "Very high" is clearly separated from the other levels, and reports an asymmetry parameter $\hat{\tau} > \frac{1}{2}$ consistent with the right-skewed incomes. Sections 5 and 6 show that these properties translate into better partition recovery and variable selection when the structure is less favorable.

5. Simulation Study

5.1. Software and Computing Environment

ExSC-$\tau\theta$ is implemented in `R` (version current in August 2026), run in RStudio 2026.07.1, with the computationally intensive parts (pairwise dissimilarities and dispersions) written in `C++` through `Rcpp`. The competitors are sparse k-means from the `sparcl` package (version 1.0.4, CRAN), `VarSelLCM` (version 2.1.3.1, CRAN), and DBI-SC, run with the authors' public code accompanying [15] (https://github.com/ffair/DBI-SC, accessed on 4 June 2026). Evaluation used `mclust` (version 6.1.1, CRAN; adjusted Rand index), `cluster` and `fpc` (silhouette, Calinski–Harabasz and Dunn indices), and `factoextra` and `ggplot2` (PCA figures). In the k-means step, nominal variables are dummy-coded. All experiments ran on a laptop with an Intel Core i7-10750H processor (10th generation, 6 cores/12 threads), 16 GB DDR4 RAM, under Windows 10, single-threaded. The random seed was fixed with `set.seed(123)` and identical datasets were supplied to all methods; each method used 100 random starts. Timing in Table 15 is wall-clock time including score estimation and tuning of τ, and excluding the grid search for K.

5.2. Simulation Setup

We conduct a series of simulation experiments to evaluate the finite-sample performance of the proposed ExSC-$\tau\theta$ method for clustering mixed-type high-dimensional data. The simulation design is intended to assess both the clustering accuracy and the ability of the proposed procedure to identify the informative variables among a large number of noise variables. Particular attention is given to the two adaptive components of ExSC-$\tau\theta$: the expectile-based dissimilarity for continuous variables and the data-driven Azzalini encoding for ordinal variables.

We consider $K \in \{2, 3, 4\}$ clusters. For a given value of K, the clusters are balanced, with

$$n_1 = n_2 = \cdots = n_K = 100, \tag{47}$$

and consequently $n = 100\,K$. Each observation contains continuous, ordinal, and nominal variables. We denote the numbers of continuous, ordinal, and nominal variables by p_c, p_o, and p_n, respectively. The simulation considers a high-dimensional mixed-data setting in which only a subset of the variables carries information about the underlying cluster structure.

Two distributional scenarios are considered. They are deliberately chosen to distinguish settings in which conventional symmetric dissimilarities are adequate from settings in which distributional heterogeneity plays an important role.

5.2.1. Scenario S1: Scale Heterogeneity

Under S1, the continuous signal variables are generated from cluster-specific distributions with different scales. The cluster separation is therefore driven primarily by differences in dispersion rather than by differences in the conditional means. The scale configurations vary with K, allowing for the difficulty of the clustering problem to increase as the number of clusters increases.

5.2.2. Scenario S2: Skewness Heterogeneity

Under S2, the continuous signal variables are generated so that the clusters exhibit different degrees and directions of skewness.

5.2.3. Ordinal and Nominal Variables

For the ordinal variables, the signal variables have cluster-dependent probabilities over ordered categories, whereas the noise ordinal variables are generated independently of the cluster labels. The nominal signal variables are similarly generated using cluster-dependent category probabilities, while the nominal noise variables contain no cluster information. The same simulated datasets are supplied to every competing method.

For every combination of K and simulation scenario, we generate $B = 50$ independent replications. All methods are evaluated using the known generating cluster labels and the known set of informative variables.

5.3. Methods for Comparison

We compare ExSC-$\tau\theta$ with three representative approaches that perform variable selection: DBI-SC, VarSelLCM, and sparse K-means (sKmeans). DBI-SC [15] is a sparse clustering method for mixed-type data in which ordinal and nominal variables are both processed through a co-occurrence distance (ordinal variables are thus treated as nominal); it provides the most direct sparse benchmark. VarSelLCM represents a model-based approach that simultaneously addresses clustering and variable selection. sKmeans is included as a sparse distance-based clustering benchmark.

The ordinal-aware method of [12], which extends k-prototypes to ordinal variables, is not included as a competitor: it performs no variable selection, so the criteria TPR/TNR of Tables 9–11 and the number of selected ordinal variables of Table 17 are undefined for it, and a comparison on clustering indices alone would confound the effect of the ordinal encoding with the effect of sparsity. The contribution of the learned ordinal encoding within our framework is isolated by the ablation study of Section 5.6, and the claims on ordinal recovery are restricted to the compared sparse methods. For the same reason, PAM with Gower's dissimilarity and k-prototypes with plain encodings, which perform no variable selection, are not included.

Table 9. Simulation results for $K = 2$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	ExSC-$\tau\theta$	0.996	0.993	0.002	0.712	0.962	1.000	0.994	0.946	0.900
S1	DBI-SC	1.000	1.000	0.000	0.840	0.920	0.080	0.979	0.942	0.977
S1	VarSelLCM	0.450	0.455	0.275	0.447	0.973	0.447	0.980	0.449	0.973
S1	sKmeans	0.698	0.617	0.151	0.998	0.681	1.000	0.677	0.992	0.681
S2	ExSC-$\tau\theta$	1.000	1.000	0.000	1.000	0.958	0.998	0.994	0.940	0.958
S2	DBI-SC	1.000	1.000	0.000	0.974	0.992	0.058	0.992	0.940	0.992
S2	VarSelLCM	0.558	0.562	0.221	0.560	0.981	0.556	0.981	0.564	0.981
S2	sKmeans	1.000	1.000	0.000	1.000	0.983	0.996	0.983	0.724	0.986

Table 10. Simulation results for $K = 3$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	ExSC-$\tau\theta$	0.986	0.977	0.006	0.996	0.996	1.000	1.000	0.938	1.000
S1	DBI-SC	0.984	0.975	0.007	0.882	1.000	0.000	1.000	0.912	1.000
S1	VarSelLCM	0.940	0.940	0.040	0.940	1.000	0.940	0.998	0.940	1.000
S1	sKmeans	0.738	0.702	0.117	0.416	0.716	1.000	0.761	0.974	0.775
S2	ExSC-$\tau\theta$	0.982	0.975	0.008	0.996	0.997	1.000	0.997	0.954	0.980
S2	DBI-SC	0.444	0.578	0.278	0.984	0.993	1.000	0.616	1.000	0.600
S2	VarSelLCM	0.405	0.458	0.346	0.400	0.999	0.400	1.000	0.400	1.000
S2	sKmeans	0.473	0.501	0.257	0.998	0.052	1.000	0.059	1.000	0.067

Table 11. Simulation results for $K = 4$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	ExSC-$\tau\theta$	0.915	0.901	0.032	0.986	0.997	1.000	0.997	0.986	0.683
S1	DBI-SC	0.799	0.842	0.083	0.906	1.000	0.422	1.000	0.808	1.000
S1	VarSelLCM	0.796	0.808	0.139	0.880	1.000	0.840	1.000	0.840	1.000
S1	sKmeans	0.572	0.617	0.167	1.000	0.697	1.000	0.697	0.894	0.694
S2	ExSC-$\tau\theta$	0.999	0.999	0.000	1.000	0.997	1.000	0.997	0.922	1.000
S2	DBI-SC	0.997	0.998	0.001	0.906	0.953	0.790	0.953	0.790	0.959
S2	VarSelLCM	0.956	0.960	0.018	0.880	1.000	0.880	1.000	0.880	1.000
S2	sKmeans	1.000	1.000	0.000	1.000	0.668	0.668	1.000	0.000	1.000

The comparison is performed on exactly the same simulated datasets. No method is given access to information about the true cluster membership or the true signal variables. The objective is therefore to compare the methods under identical experimental conditions.

5.4. Evaluation Criteria

The methods are assessed from three complementary perspectives: clustering accuracy against the known labels, variable-selection accuracy, and internal partition quality.

5.4.1. External Criteria (Known Labels)

We report the adjusted Rand index (ARI) [32], the normalized mutual information (NMI) with geometric normalization $I(A;B)/\sqrt{H(A)H(B)}$ [33], and the clustering error rate (CER), i.e., the minimum misclassification rate over label permutations. Higher ARI and NMI indicate better agreement between the estimated and true partitions, whereas a smaller CER indicates better clustering.

5.4.2. Variable Selection (Simulations)

Variable-selection performance is evaluated separately for the three types of variables. We report TPR_c, TNR_c, TPR_o, TNR_o, TPR_n, and TNR_n, where the subscripts c, o and n refer to continuous, ordinal and nominal variables, respectively. With $\hat{S}$ the set of selected variables and S^* the set of informative ones, $TPR = |\hat{S} \cap S^*|/|S^*|$ is the proportion of informative variables correctly selected and $TNR = |\hat{S}^c \cap S^{*c}|/|S^{*c}|$ the proportion of noise variables correctly excluded. A successful sparse clustering method should maintain both high TPR and high TNR.

5.4.3. Internal Criteria (Real Data)

When the labels are not used, partition quality is assessed by indices that combine within-cluster compactness and between-cluster separation, computed on the Gower dissimilarity of all variables, which is common to all methods: the Davies–Bouldin index [30] (lower is better), the average silhouette width [34] (cohesion and separation, in $[-1, 1]$), the Calinski–Harabasz index [35] (ratio of between- to within-cluster dispersion; higher is

better) and the Dunn index [36] (minimum inter-cluster distance over maximum cluster diameter; higher is better). The cophenetic correlation is defined for dendrograms of hierarchical methods and is therefore not applicable to the partitional methods compared here. Internal indices computed on all variables do not reward the removal of noise variables; they complement, and do not replace, the external indices.

All reported simulation results are averaged over the 50 independent replications.

5.5. Simulation Results

The main simulation results for $K = 2$, $K = 3$, and $K = 4$ are reported in Tables 9, 10, and 11, respectively. The discussion focuses on the joint behavior of clustering accuracy and variable-selection accuracy, rather than on ARI alone.

5.5.1. Results for $K = 2$

For two clusters, the clustering task is generally well separated. Under S1, ExSC-$\tau\theta$ obtains ARI = 0.996 and NMI = 0.993, whereas DBI-SC obtains ARI = 1.000 and NMI = 1.000. Hence, in this setting, DBI-SC has a small advantage in terms of partition recovery. However, the variable-selection results reveal a substantially different picture. ExSC-$\tau\theta$ identifies all ordinal signal variables, with $\mathrm{TPR}_o = 1.000$, whereas DBI-SC obtains only 0.080. The proposed method also retains high nominal sensitivity, with $\mathrm{TPR}_n = 0.946$.

Under S2, the four methods achieve very high clustering accuracy. ExSC-$\tau\theta$ reaches ARI and NMI equal to one. More importantly, it obtains $\mathrm{TPR}_c = 1.000$ and $\mathrm{TPR}_o = 0.998$, while maintaining high TNR values. DBI-SC also recovers the cluster partition almost perfectly, but its ordinal TPR is only 0.058. These results show that near-perfect clustering does not necessarily imply successful variable selection.

5.5.2. Results for $K = 3$

The three-cluster setting provides a more demanding assessment of the methods. Under S1, ExSC-$\tau\theta$ and DBI-SC have very similar clustering performance, with ARI values of 0.986 and 0.984, respectively, and NMI values of 0.977 and 0.975. Thus, the two methods provide essentially comparable partition recovery. The difference becomes clear in variable selection. ExSC-$\tau\theta$ obtains $\mathrm{TPR}_c = 0.996$ and $\mathrm{TPR}_o = 1.000$, whereas DBI-SC obtains $\mathrm{TPR}_c = 0.882$ and $\mathrm{TPR}_o = 0.000$.

Under S2, the advantage of the proposed method becomes much more pronounced. ExSC-$\tau\theta$ achieves ARI = 0.982, NMI = 0.975, and CER = 0.008. In contrast, DBI-SC obtains ARI = 0.444, NMI = 0.578, and CER = 0.278. The improvement is accompanied by strong variable-selection performance, with $\mathrm{TPR}_c = 0.996$ and $\mathrm{TPR}_o = 1.000$ for ExSC-$\tau\theta$.

These results suggest that the expectile component becomes particularly useful when the continuous variables exhibit heterogeneous skewness. Importantly, the gain is not obtained by simply selecting more variables: ExSC-$\tau\theta$ maintains high TNR values while achieving high sensitivity.

5.5.3. Results for $K = 4$

The four-cluster setting is the most demanding of the considered experiments. Under S1, ExSC-$\tau\theta$ obtains ARI = 0.915 and NMI = 0.901, compared with ARI = 0.799 and NMI = 0.842 for DBI-SC. The corresponding CER values are 0.032 and 0.083. Thus, the proposed method provides a clear improvement in partition recovery when the cluster structure becomes more complex.

The variable-selection results are also favorable. ExSC-$\tau\theta$ achieves TPRs of 0.986, 1.000, and 0.986 for continuous, ordinal, and nominal variables, respectively. DBI-SC obtains 0.906, 0.422, and 0.808. The difference is particularly evident for the ordinal variables.

Under S2, clustering is easier for all methods. ExSC-$\tau\theta$ achieves ARI = 0.999 and NMI = 0.999, while DBI-SC obtains 0.997 and 0.998. Although the difference in clustering accuracy is small, ExSC-$\tau\theta$ maintains strong variable-selection performance, including perfect TPR for both continuous and ordinal variables.

Overall, the main experiments indicate that the advantage of ExSC-$\tau\theta$ is most evident in the more challenging settings. When the cluster partition is already easy to recover, several methods obtain nearly perfect ARI values. In contrast, the differences become substantial when distributional heterogeneity and multiple clusters make the underlying geometry more difficult.

5.6. Ablation Study

To investigate the contribution of the two adaptive components of ExSC-$\tau\theta$, we perform an ablation study for each value of K. The first component is the expectile parameter τ, which controls the asymmetric treatment of continuous-variable differences. The second component is the Azzalini encoding parameter θ_j, which determines the learned numerical representation of ordinal variables.

We consider four variants. A0 corresponds to the ExSC-$\tau\theta$ framework with both adaptive components disabled: the quantile parameter is fixed at $\tau = \frac{1}{2}$, and the ordinal scores are fixed using the Johnson S_U family with $\theta_j = (0, 1)$. Under these settings, ExSC-$\tau\theta$ reduces to a DBI-SC framework in terms of its clustering objective, while retaining a difference in the treatment of ordinal variables. Specifically, ExSC-$\tau\theta$ transforms ordinal categories into metric variables using fixed Azzalini-based scores and incorporates them into the metric dissimilarity component. In contrast, classical DBI-SC, following [15], treats ordinal variables as nominal and incorporates them through the co-occurrence dissimilarity. This distinction explains why A0 and classical DBI-SC may produce identical or similar results in some cases, while their results can also differ. Thus, A0 serves as a fixed-parameter variant of ExSC-$\tau\theta$ rather than an exact reproduction of classical DBI-SC. A1 estimates τ while keeping the ordinal scoring parameters fixed at $\theta_j = (0, 1)$. A2 estimates θ_j while fixing $\tau = \frac{1}{2}$. Finally, A3 estimates both τ and θ_j and therefore corresponds to the complete ExSC-$\tau\theta$ procedure.

5.6.1. Ablation Results for $K = 2$

Table 12 reports the ablation results for $K = 2$.

Table 12. Ablation study for $K = 2$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	A0	0.993	0.987	0.004	0.002	1.000	0.080	0.979	0.942	0.977
S1	A1: τ learned	0.995	0.991	0.002	0.732	0.961	0.030	0.994	0.946	0.900
S1	A2: θ learned	0.992	0.986	0.004	0.002	1.000	1.000	1.000	0.928	0.940
S1	A3: ExSC-$\tau\theta$	0.996	0.993	0.002	0.712	0.962	1.000	0.994	0.946	0.900
S2	A0	1.000	1.000	0.000	0.000	0.974	0.058	0.994	0.992	0.992
S2	A1: τ learned	1.000	1.000	0.000	0.996	0.958	0.053	0.823	0.942	0.920
S2	A2: θ learned	1.000	1.000	0.000	0.000	0.973	0.995	0.999	0.942	0.920
S2	A3: ExSC-$\tau\theta$	1.000	1.000	0.000	1.000	0.958	0.998	0.994	0.940	0.958

For $K = 2$, the ablation results illustrate the complementary roles of the two components. Under S1, learning τ substantially improves continuous-variable recovery, while learning θ_j is responsible for recovering the ordinal signal. A1 has $TPR_c = 0.732$ but only $TPR_o = 0.030$, whereas A2 has $TPR_o = 1.000$ but $TPR_c = 0.002$. The complete method combines these two properties and obtains $TPR_o = 1.000$ together with a substantially improved continuous TPR.

Under S2, the same pattern is observed more clearly. A1 achieves $TPR_c = 0.996$ but $TPR_o = 0.053$, whereas A2 achieves $TPR_o = 0.995$ but no continuous recovery. The complete method simultaneously reaches $TPR_c = 1.000$ and $TPR_o = 0.998$.

5.6.2. Ablation Results for $K = 3$

Table 13 reports the ablation results for $K = 3$.

Table 13. Ablation study for $K = 3$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	A0	0.985	0.977	0.006	0.000	0.500	0.000	0.360	0.942	1.000
S1	A1: τ learned	0.987	0.979	0.006	0.996	0.996	0.010	0.040	0.938	1.000
S1	A2: θ learned	0.985	0.977	0.006	0.000	0.700	0.000	0.750	0.942	1.000
S1	A3: ExSC-$\tau\theta$	0.986	0.977	0.006	0.996	0.996	1.000	0.996	0.938	1.000
S2	A0	0.444	0.578	0.278	0.000	0.616	0.870	0.634	0.570	0.650
S2	A1: τ learned	0.789	0.774	0.108	0.998	0.899	0.600	0.767	0.470	0.550
S2	A2: θ learned	0.639	0.564	0.208	0.000	0.777	0.930	0.976	0.870	0.813
S2	A3: ExSC-$\tau\theta$	0.982	0.975	0.008	0.996	0.997	1.000	0.997	0.954	0.980

The $K = 3$ experiments provide stronger evidence for the interaction between the two adaptive components. Under S1, A1 recovers the continuous signal almost completely but fails to recover the ordinal signal. Conversely, A2 does not adequately recover the continuous signal. The complete procedure achieves high sensitivity for both variable types while retaining high specificity.

The difference is particularly pronounced under S2. A0 has ARI = 0.444, while A1 and A2 reach 0.789 and 0.639, respectively. The complete method reaches ARI = 0.982 and NMI = 0.975, with CER = 0.008. This result indicates that the two adaptive components provide complementary information rather than representing interchangeable modifications of the same procedure.

5.6.3. Ablation Results for $K = 4$

Table 14 reports the ablation results for $K = 4$.

Table 14. Ablation study for $K = 4$. Values are averaged over 50 replications.

Scenario	Method	ARI	NMI	CER	TPR_c	TNR_c	TPR_o	TNR_o	TPR_n	TNR_n
S1	A0	0.799	0.842	0.083	0.000	0.501	1.000	0.004	0.808	1.000
S1	A1: τ learned	0.920	0.910	0.028	0.990	0.996	0.025	0.002	0.982	0.665
S1	A2: θ learned	0.957	0.949	0.016	0.000	0.403	0.900	0.922	0.902	0.555
S1	A3: ExSC-$\tau\theta$	0.915	0.901	0.032	0.986	0.997	1.000	0.997	0.986	0.683
S2	A0	0.957	0.948	0.016	0.000	0.953	0.790	0.953	0.790	0.959
S2	A1: τ learned	0.925	0.910	0.028	1.000	0.997	0.800	0.955	0.976	0.876
S2	A2: θ learned	0.957	0.949	0.016	0.000	0.768	1.000	0.999	0.900	0.999
S2	A3: ExSC-$\tau\theta$	0.999	0.999	0.000	1.000	0.997	1.000	0.997	0.922	1.000

For four clusters, the roles of the two adaptive components remain distinct. Under S1, A1 obtains $TPR_c = 0.990$ but only $TPR_o = 0.025$, whereas A2 obtains $TPR_o = 0.900$ but $TPR_c = 0.000$. The complete method simultaneously reaches $TPR_c = 0.986$ and $TPR_o = 1.000$.

Under S2, the full procedure provides the strongest overall result, with ARI and NMI both equal to 0.999 and CER equal to zero. It also achieves perfect TPR for both continuous and ordinal variables. These results support the interpretation that the expectile component and the learned ordinal encoding address different sources of heterogeneity in mixed-type data.

Across all three values of K, the ablation experiments therefore provide empirical evidence that neither adaptive component alone is sufficient to reproduce the behavior of

the complete method. Learning τ primarily affects continuous-variable recovery, whereas learning θ_j is essential for exploiting ordinal information. Their joint estimation provides a more balanced treatment of the three variable types.

5.7. Behavior of the Estimated Expectile Parameter

The estimated expectile parameter provides an additional diagnostic of the continuous-variable geometry. The value $\tau = 0.5$ corresponds to a symmetric treatment of deviations, while values away from 0.5 introduce asymmetric weighting.

The estimated values obtained in the simulation experiment are different from 0.5 in the considered settings. For example, the complete procedure gives $\widehat{\tau} = 0.547$ for $K = 2$ under one setting and $\widehat{\tau} = 0.594$ under the corresponding asymmetric setting, while larger deviations from 0.5 are observed as the cluster structure becomes more heterogeneous. At $K = 4$, the estimated value reaches approximately $\widehat{\tau} = 0.641$.

These estimates should not be interpreted as evidence that one particular value of τ is universally optimal. Rather, they illustrate that the proposed procedure adapts the continuous dissimilarity to the empirical distribution of the data. The ablation results support this interpretation: fixing $\tau = 0.5$ can substantially reduce continuous-variable recovery in settings with heterogeneous distributional shapes.

5.8. Computational Time Comparison

In addition to statistical performance, we compare the computational cost of the four methods. The mean execution time and its standard deviation over the 50 replications are reported in Table 15. The comparison is performed under the same computational environment and using the same simulated datasets.

As expected, sKmeans has the smallest execution time in every setting. This computational advantage, however, should be interpreted together with its substantially different treatment of mixed-type variables. ExSC-$\tau\theta$ requires additional computation because it estimates the expectile parameter and the ordinal encoding, but its computational cost remains considerably below that of DBI-SC.

Table 15. Mean (SD) execution time in seconds over 50 replications.

Scenario	K	ExSC-$\tau\theta$	DBI-SC	VarSelLCM	sKmeans
S1	2	2.22 (0.41)	12.26 (1.64)	3.44 (0.99)	0.06 (0.02)
S1	3	4.53 (0.65)	65.56 (14.38)	7.25 (5.70)	0.17 (0.07)
S1	4	7.47 (0.87)	134.07 (26.03)	9.56 (0.57)	0.27 (0.10)
S2	2	1.77 (0.06)	8.93 (0.50)	2.14 (0.32)	0.03 (0.01)
S2	3	3.44 (0.66)	43.32 (5.73)	5.94 (3.14)	0.07 (0.02)
S2	4	4.32 (0.47)	71.45 (18.30)	13.91 (11.77)	0.10 (0.01)

The difference becomes increasingly important as the number of clusters increases. Under S1, the mean execution time of ExSC-$\tau\theta$ increases from 2.22 s for $K = 2$ to 7.47 s for $K = 4$, whereas DBI-SC increases from 12.26 to 134.07 s. Under S2, the corresponding times for ExSC-$\tau\theta$ are 1.77, 3.44, and 4.32 s, compared with 8.93, 43.32, and 71.45 s for DBI-SC. Thus, the proposed method provides a favorable compromise between the additional flexibility of adaptive dissimilarities and computational cost.

5.9. Discussion of the Simulation Results

Across the three values of K, the two simulation scenarios, and the four competing procedures, the results lead to several consistent conclusions.

Clustering accuracy and variable-selection accuracy should be considered jointly. The simulations show that a method may recover the cluster partition accurately while failing

to identify the variables responsible for the separation. This is particularly clear for $K = 2$. Under S1, DBI-SC achieves ARI = 1.000, compared with 0.996 for ExSC-$\tau\theta$, but its ordinal TPR is only 0.080, whereas the proposed method achieves 1.000. Thus, a comparison based only on ARI would obscure an important difference in the scientific interpretation of the clusters.

The learned ordinal representation is a major source of the improvement. Across the simulations, ExSC-$\tau\theta$ consistently obtains very high ordinal TPR. The contrast is especially pronounced for $K = 3$ and $K = 4$. Under S1 with $K = 3$, DBI-SC obtains $\text{TPR}_o = 0.000$, whereas ExSC-$\tau\theta$ obtains 1.000. For $K = 4$, the corresponding values under S1 are 0.422 and 1.000. These results indicate that treating ordinal variables through a learned encoding can be important when their category spacings are related to the underlying cluster structure. Among the compared sparse methods, only ExSC-$\tau\theta$ attains an ordinal TPR of at least 0.998 in every main experiment.

The expectile component becomes most useful under distributional heterogeneity. The clearest example occurs under S2 with $K = 3$, where ExSC-$\tau\theta$ achieves ARI = 0.982, while DBI-SC achieves only 0.444. A similar improvement is observed under S1 with $K = 4$, where the ARI increases from 0.799 for DBI-SC to 0.915 for ExSC-$\tau\theta$. These findings are consistent with the motivation for the expectile dissimilarity: the method does not require all continuous-variable differences to contribute symmetrically to the clustering criterion.

The ablation study confirms that the two adaptive mechanisms are complementary. Learning τ primarily improves continuous-variable recovery, whereas learning θ_j primarily improves ordinal-variable recovery. At $K = 3$ under S2, for example, A1 and A2 obtain ARI values of 0.789 and 0.639, respectively, while the complete method reaches 0.982. At $K = 4$ under S2, the complete procedure reaches ARI and NMI of 0.999 and obtains perfect TPR for both continuous and ordinal variables. The evidence therefore supports the joint formulation rather than either adaptive component in isolation.

The computational results indicate that the additional flexibility remains practical. sKmeans is consistently the fastest method, but it does not provide the same treatment of the different variable types. ExSC-$\tau\theta$ is more computationally demanding than sKmeans because of the additional parameter estimation, yet it remains substantially faster than DBI-SC. At $K = 4$, for example, ExSC-$\tau\theta$ requires 7.47 s under S1 compared with 134.07 s for DBI-SC, and 4.32 s under S2 compared with 71.45 s.

Taken together, the simulations do not suggest that ExSC-$\tau\theta$ dominates every competing method on every individual metric. Rather, they demonstrate that its main advantage lies in providing a balanced framework for mixed-type sparse clustering. The method can adapt the continuous-variable dissimilarity to distributional asymmetry, learn a meaningful representation for ordinal variables, and maintain high variable-selection sensitivity and specificity. The gains are most pronounced in settings where the underlying cluster geometry is not adequately captured by a fixed symmetric distance.

6. Real Data Application

We apply ExSC-$\tau\theta$ to two real benchmark datasets from different sectors: Steel Annealing (manufacturing, Section 6.1), which jointly covers continuous physical measurements, genuinely ordered categorical levels with unknown spacing, and nominal variables with structured inter-category dependencies, and Credit Approval (banking, Section 6.2), a continuous/nominal mixture whose single ordinal variable is constructed. The datasets carry external class labels that are withheld during clustering and used only for evaluation via external criteria (ARI, NMI) and internal criteria computed on the Gower dissimilarity (Davies–Bouldin, silhouette, Calinski–Harabasz and Dunn indices; Section 5.4).

6.1. Dataset: Steel Annealing

6.1.1. Data Description

The Steel Annealing dataset [37] is drawn from the UCI Machine Learning Repository and contains $n = 798$ observations of steel samples described by physical and chemical properties measured before and after the annealing process. After removing constant columns and incomplete cases, the feature matrix comprises

- Six continuous variables: carbon, hardness, strength, thick, width, and len. These measurements describe the physical and chemical composition of the steel, which may be asymmetrically distributed, so that an expectile dissimilarity with $\hat{\tau} \neq \frac{1}{2}$ is potentially of interest compared with the symmetric squared-Euclidean distance.
- Four ordinal variables: formability (six ordered levels), enamelability (six ordered levels), surface_quality (five ordered levels: D, E, F, G, UNKNOWN), and packing (four ordered levels). The level spacings are determined by the physical process and are not equidistant a priori, making data-driven Azzalini scores directly relevant [12].
- Twenty-one nominal variables: including family, product_type, steel, condition, non_ageing, surface_finish, and various binary chemical indicators (bc, bf, bt, bw_me, bl, m, chrom, phos, cbond, marvi, exptl, ferro, corr, lustre, and shape). These unordered categories carry structured co-occurrence dependencies that the Ahmad–Dey dissimilarity is designed to capture.

The class variable takes the values $\{1, 2, 3, 5, U\}$ (class 4 does not occur in the 798 instances), so there are $K^* = 5$ non-empty classes, of sizes 8, 88, 608, 60 and 34 respectively; the data are strongly imbalanced (class 3 represents 76.2% of the observations). The class labels are withheld during clustering. The number of clusters is selected by minimizing the Davies–Bouldin index over $K \in \{2, \ldots, 12\}$, which gives $K = 5$, equal to the number of non-empty classes (Figure 1).

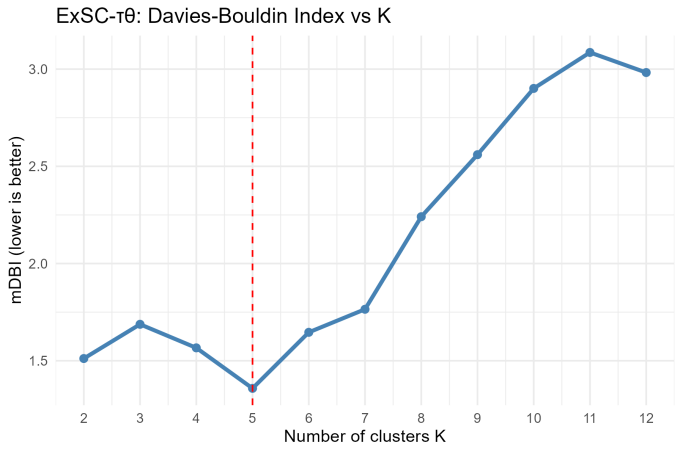


Figure 1. Davies–Bouldin index of ExSC-$\tau\theta$ as a function of the number of clusters K on the Steel Annealing dataset ($n = 798$). The dashed red line marks the optimal $K^* = 5$ selected by the minimum DBI criterion, which equals the number of non-empty classes.

6.1.2. Preprocessing

Continuous variables are standardized to zero mean and unit variance. Ordinal variables are encoded as ordered factors with their natural level ordering; the (UNKNOWN) category is included as the lowest level where it appears. All ordinal integer codes are then passed to the Azzalini score estimator, which learns the actual inter-level spacing from the data. Nominal variables are converted to integer-coded factors before computing the co-occurrence dissimilarity.

6.1.3. Selection of the Number of Clusters

Figure 1 displays the Davies–Bouldin index for ExSC-$\tau\theta$ as a function of K. The curve exhibits a clear minimum at $K = 5$, which equals the number of non-empty classes. All four methods are then evaluated at this common $K = 5$.

6.1.4. Results

Table 17 reports external and internal indices for each method at $K = 5$. Among the four variable-selecting methods, ExSC-$\tau\theta$ attains the best ARI (0.247), NMI (0.378), Davies–Bouldin index (1.358), silhouette width (0.326) and Dunn index (0.117). The Calinski–Harabasz index of DBI-SC is marginally higher (195.93 versus 195.09), so the two methods are essentially tied on this criterion. DBI-SC attains ARI = 0.054 and DBI = 2.014; sKmeans ARI = 0.013 and DBI = 2.934; VarSelLCM ARI = 0.005 and DBI = 3.689. In absolute terms, ARI = 0.247 indicates only partial agreement with the class labels, which is expected since the classes are strongly imbalanced and only partly determined by the recorded variables. The estimated asymmetry parameter $\hat{\tau} = 0.891$ is well above $\frac{1}{2}$: positive pairwise gaps receive a weight $2\hat{\tau} = 1.782$ and negative gaps $2(1 - \hat{\tau}) = 0.218$.

The ordinal-variable selection results illustrate the behavior of the method on ordinal data. ExSC-$\tau\theta$ assigns nonzero weight to two out of four ordinal variables (*formability* and *surface_quality*), whose Azzalini scores learn spacings tailored to the provisional cluster structure, whereas integer coding would impose equal spacing. DBI-SC processes all ordinal variables through the co-occurrence distance designed for unordered nominal categories, thus discarding their ordering, and assigns nonzero weight to four out of four ordinal variables. sKmeans assigns nonzero weight to two out of four ordinal variables with equal-spacing integer codes. VarSelLCM selects no ordinal variable.

6.1.5. PCA Visualization

Figure 2 shows the projections of all $n = 798$ observations onto the first two principal components (PC1 (41.2%) and PC2 (11.1%) of the variance), with the cluster assignments of the four methods. PCA is used here for visualization only; clustering is performed in the original variable space. As the two components explain 52.3% of the total variance, overlap or separation in the plane is suggestive but not conclusive, and the quantitative assessment is that of Table 17. The ExSC-$\tau\theta$ partition (top left) shows the smallest ellipse overlap among the four methods. DBI-SC (top right) and sKmeans (bottom left) show more overlap between adjacent clusters, and VarSelLCM (bottom right) is dominated by one large cluster (591 observations), consistent with its ARI close to zero.

6.1.6. Interpretation of the Clusters

Table 16 cross-tabulates the five ExSC-$\tau\theta$ clusters (of sizes 507, 48, 123, 57 and 63) with the true classes. Cluster 5 is almost pure class 5 (58 of its 60 members); cluster 3 contains all 34 instances of class U together with 88 of class 3; cluster 1 is large and gathers 71 of the 88 class-2 instances and most of class 3; cluster 2 (48 observations) is a pure subgroup of class 3; and cluster 4 collects seven of the eight class-1 instances. Majority-class purity is $661/798 = 0.828$ for ExSC-$\tau\theta$, versus $608/798 = 0.762$ (the share of the largest class) for each of the other three methods. A pure 48-observation subgroup and a group of 57 observations (class composition 7/17/33) also appear in the DBI-SC and sKmeans partitions, so the methods differ mainly in how the remaining observations are split. The ordinal variables with nonzero weight are formability and surface_quality, while enamelability and packing receive zero weight.

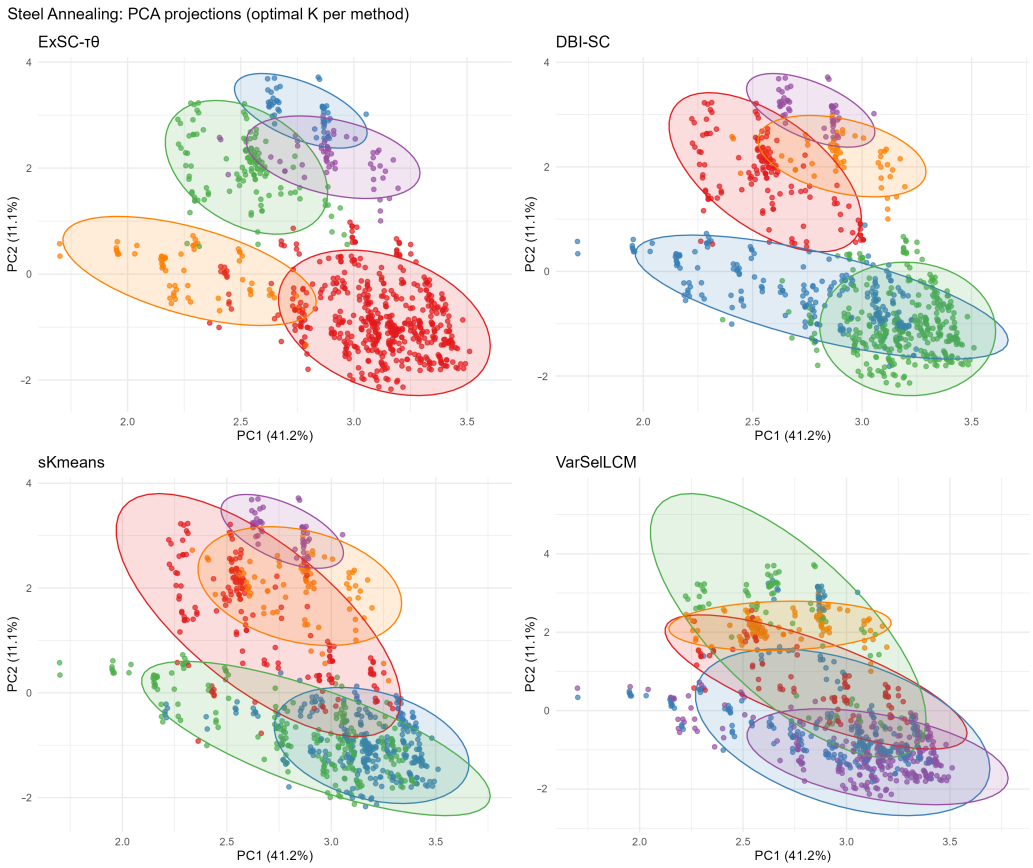


Figure 2. PCA projections (PC1 vs. PC2) of the Steel Annealing dataset ($n = 798$, $K = 5$) for all four clustering methods. Ellipses are 95% confidence ellipses for each cluster. ExSC-$\tau\theta$ (**top left**) shows the smallest ellipse overlap; VarSelLCM (**bottom right**) is dominated by one large cluster.

Table 16. Contingency table of the ExSC-$\tau\theta$ clusters versus the true classes (Steel Annealing, $K = 5$).

Cluster	1	2	3	5	U
1	0	71	434	2	0
2	0	0	48	0	0
3	1	0	88	0	34
4	7	17	33	0	0
5	0	0	5	58	0

6.2. Dataset: Credit Approval

6.2.1. Data Description

The Credit Approval dataset [38] is drawn from the UCI Machine Learning Repository and concerns credit card applications. All attribute names and values are anonymized for confidentiality. It contains $n = 690$ applications described by 15 attributes. The variables are as follows:

- Five continuous variables: A2, A3, A8, A14 and A15.
- One ordinal variable: A11, a count variable (from 0 to 67) that we discretize into five ordered classes {0, 1, 2–3, 4–6, 7^+}.
- Nine nominal variables: A1, A4, A5, A6, A7, A9, A10, A12 and A13.

6.2.2. Preprocessing and Selection of the Number of Clusters

Continuous variables are standardized to zero mean and unit variance; the ordinal variable is an ordered factor passed to the Azzalini score estimator; nominal variables are integer-coded factors. As for Steel Annealing, K is selected by minimizing the modified

Davies–Bouldin index of ExSC-$\tau\theta$ over $K \in \{2, \ldots, 12\}$ (Figure 3). The minimum is reached at $K = 3$, and the four methods are then evaluated at this common value.

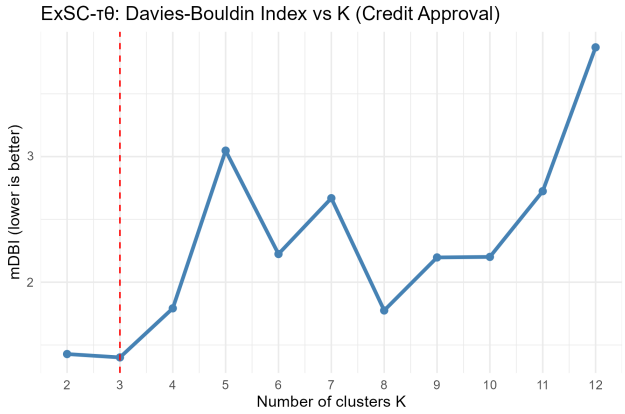


Figure 3. Modified Davies–Bouldin index of ExSC-$\tau\theta$ as a function of the number of clusters K on the Credit Approval dataset ($n = 690$). The dashed red line marks the minimum, $K = 3$.

6.2.3. PCA Visualization

Figure 4 shows the projections of all $n = 690$ observations onto the first two principal components (PC1 (20.2%) and PC2 (13.3%) of the variance), with the cluster assignments of the four methods. PCA is used for visualization only; clustering is performed in the original variable space. As the two components explain only 33.5% of the total variance, overlap or separation in the plane is suggestive but not conclusive, and the quantitative assessment is that of Table 18. The ExSC-$\tau\theta$ clusters appear more separated along PC1 than those of the other methods, which overlap strongly.

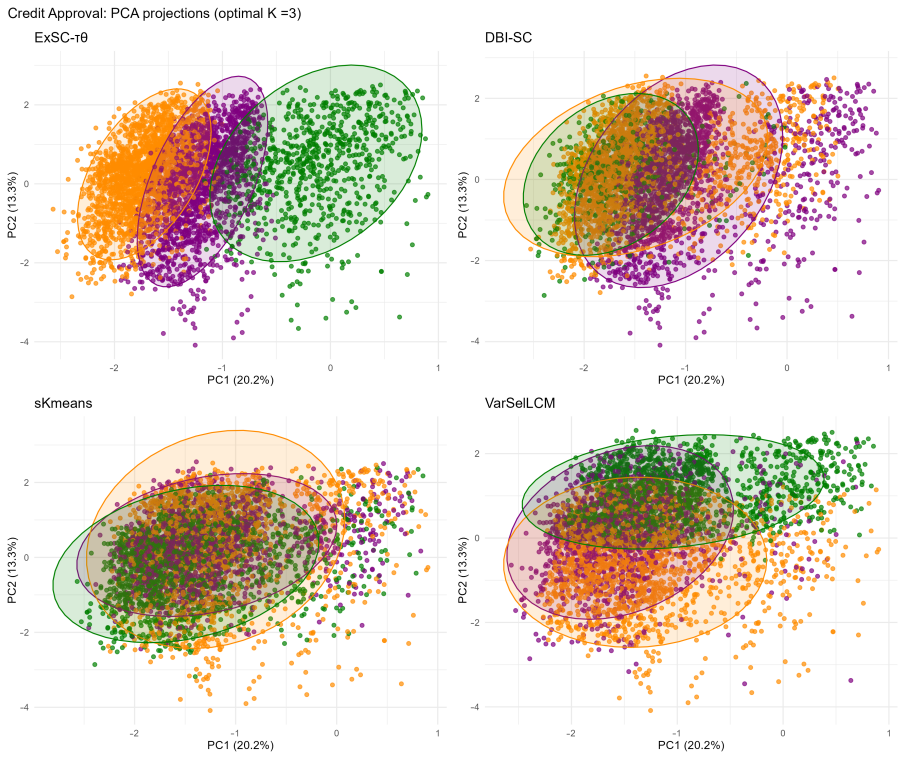


Figure 4. PCA projections (PC1 vs. PC2) of the Credit Approval dataset ($n = 690$, $K = 3$) for the four clustering methods. Ellipses are 95% confidence ellipses for each cluster.

6.3. Summary of the Results

6.3.1. Steel Annealing

Table 17 presents the results for Steel Annealing at $K = 5$. ExSC-$\tau\theta$ achieves the highest ARI (0.247) and NMI (0.378), together with the lowest Davies–Bouldin index (1.358), the highest silhouette width (0.326), and the highest Dunn index (0.117). Its ARI exceeds that of DBI-SC, sKmeans, and VarSelLCM by 0.193, 0.234, and 0.242, respectively. The Calinski–Harabasz indices of ExSC-$\tau\theta$ and DBI-SC are close, at 195.09 and 195.93, respectively. Nevertheless, the ARI of 0.247 indicates that agreement with the true class labels remains moderate.

Table 17. Real data results on Steel Annealing ($n = 798$, $K = 5$). ARI and NMI are computed against the true classes. DBI, silhouette (Sil.), Calinski–Harabasz (CH), and Dunn indices are computed using the Gower dissimilarity matrix based on all variables, held common across methods. $\hat{\tau}$ is reported for ExSC-$\tau\theta$ only. Ord. > 0 denotes the number of ordinal variables with nonzero weight out of 4.Bold values indicate the best result for each evaluation criterion.

Method	ARI	NMI	DBI	Sil.	CH	Dunn	$\hat{\tau}$	Ord. > 0
ExSC-$\tau\theta$	**0.247**	**0.378**	**1.358**	**0.326**	195.09	**0.117**	0.891	2/4
DBI-SC	0.054	0.229	2.014	0.270	**195.93**	0.048	—	4/4
sKmeans	0.013	0.208	2.934	0.193	150.07	0.020	—	2/4
VarSelLCM	0.005	0.090	3.689	−0.039	53.42	0.008	—	0/4

The variable weights also provide insight into the role of ordinal features. ExSC-$\tau\theta$ assigns nonzero weights to two of the four ordinal variables, namely formability and surface_quality, while enamelability and packing receive zero weights. In contrast, DBI-SC assigns nonzero weights to all four ordinal variables. These results illustrate how the proposed method can select a subset of ordinal features rather than retaining the entire ordinal block.

6.3.2. Credit Approval

Table 18 reports the external and internal clustering indices obtained by the four methods on the Credit Approval dataset at $K = 3$, using the same evaluation protocol as for Steel Annealing. ExSC-$\tau\theta$ achieves the highest ARI (0.196) and NMI (0.149), the lowest Davies–Bouldin index (1.401), the highest silhouette width (0.266), and the highest Dunn index (0.088). DBI-SC obtains a slightly higher Calinski–Harabasz index than ExSC-$\tau\theta$ (153.72 versus 151.43). The ARI values for DBI-SC, sKmeans, and VarSelLCM are 0.164, 0.030, and 0.070, respectively.

In terms of variable selection, ExSC-$\tau\theta$ retains 14 of the 15 variables, compared with nine variables for DBI-SC and ten variables for VarSelLCM. The estimated asymmetry parameter is $\hat{\tau} = 0.731$. The single ordinal variable (the discretized count A11) receives a nonzero weight in all four methods.

Table 18. Real data results on Credit Approval ($n = 690$, $K = 3$). ARI and NMI are computed against the approval class. DBI, silhouette (Sil.), Calinski–Harabasz (CH), and Dunn indices are computed using the Gower dissimilarity matrix based on all variables, held common across methods. $\hat{\tau}$ is reported for ExSC-$\tau\theta$ only. Sel. denotes the number of variables with nonzero weight out of 15; Ord. > 0 denotes the number of ordinal variables with nonzero weight out of 1. Bold values indicate the best result for each evaluation criterion.

Method	ARI	NMI	DBI	Sil.	CH	Dunn	$\hat{\tau}$	Sel.	Ord. > 0
ExSC-$\tau\theta$	**0.196**	**0.149**	**1.401**	**0.266**	151.43	**0.088**	0.731	14/15	1/1
DBI-SC	0.164	0.141	1.627	0.194	**153.72**	0.056	—	9/15	1/1
sKmeans	0.030	0.037	3.018	0.054	33.93	0.033	—	15/15	1/1
VarSelLCM	0.070	0.077	4.847	0.005	29.27	0.005	—	10/15	1/1

6.3.3. Overall Assessment

Overall, the two real data applications provide complementary evidence on the behavior of ExSC-$\tau\theta$. On Steel Annealing, the method performs favorably in terms of external agreement, internal clustering indices, and ordinal variable selection. On Credit Approval, it also achieves the best values for most indices. The two applications should therefore be interpreted as illustrative empirical evidence rather than as a general demonstration of superiority.

7. Conclusions

7.1. Summary of the Contribution

This paper introduced ExSC-$\tau\theta$, a sparse clustering framework for high-dimensional data combining continuous, ordinal, and nominal variables. Three design choices differentiate it from existing methods: an expectile dissimilarity parameterized by $\tau \in (0, 1)$ that captures directional asymmetry in the cluster structure; a data-driven Azzalini score transformation that learns the spacing of ordinal levels from a cluster-quality criterion; and a co-occurrence dissimilarity for nominal variables, unified with the metric components through a single ℓ_1-penalized sparse optimization.

7.2. Key Findings

The simulation studies across $K \in \{2, 3, 4\}$, two distributional scenarios and 50 replications, together with the two real data applications, yielded the following main findings.

(i) *Ordinal variable recovery.* ExSC-$\tau\theta$ attains an ordinal TPR of at least 0.998 in all main experiments. No compared sparse competitor does so: DBI-SC, which processes ordinal variables through a co-occurrence distance and thus treats them as unordered, has an ordinal TPR as low as 0.000; VarSelLCM reaches at most 0.940; sKmeans attains a high ordinal TPR in several settings but with low specificity (e.g., $\mathrm{TNR}_o = 0.059$ for S2, $K = 3$) and an ordinal TPR of 0.668 for S2, $K = 4$. This supports the methodological claim that estimating θ_j by maximizing the adjusted DBI at a provisional partition learns score spacings related to the actual cluster separation.

(ii) *Estimated asymmetry.* The estimated $\hat{\tau}$ differs from $\frac{1}{2}$ in the considered settings (Section 5.7) and is returned alongside the partition as a diagnostic of asymmetry. We do not claim that a particular value of τ is universally optimal; the ablation study shows that fixing $\tau = \frac{1}{2}$ can substantially reduce continuous-variable recovery under heterogeneous distributional shapes.

(iii) *Clustering accuracy.* Gains are largest under distributional heterogeneity: ExSC-$\tau\theta$ leads DBI-SC by 0.116 ARI for scale heterogeneity at $K = 4$ and by 0.538 ARI for skewness heterogeneity at $K = 3$ (0.982 versus 0.444), while differences are negligible, or in favor of DBI-SC, where mean separation dominates ($K = 2$).

(iv) *Computational efficiency.* At $K = 4$ (S1), ExSC-$\tau\theta$ runs about 18× faster than DBI-SC (7.47 s versus 134.07 s) and is faster than VarSelLCM in all settings of Table 15; sKmeans is the fastest method. ExSC-$\tau\theta$ remains quadratic in n.

(v) *Real data.* On the Steel Annealing dataset ($n = 798$, $K = 5$ non-empty classes $\{1, 2, 3, 5, U\}$), ExSC-$\tau\theta$ attains the best ARI (0.247 versus 0.054 for DBI-SC), NMI, DBI, silhouette and Dunn index among the variable-selecting methods and is tied with DBI-SC on the Calinski–Harabasz index; it assigns nonzero weight to two of four ordinal variables. On the Credit Approval dataset ($n = 690$, $K = 3$), ExSC-$\tau\theta$ retains 14 of 15 variables, and the ordinal block is a discretized count. These two applications illustrate, rather than generalize, the simulation findings.

7.3. Theoretical Contribution

Theorem 1 establishes the screening consistency of the variable-selection step under conditions on signal strength ($\xi_1, \xi_3 > 0$), dimensionality growth ($\ln(p + r + q) = o(n)$) and sparsity calibration. The proof accommodates subexponential concentration for metric variables under Gaussian data, Bernstein-type concentration for bounded nominal dissimilarities, and an $O_p(n^{-1/2})$ plug-in perturbation from the pre-estimated ordinal scores; the role of τ is handled through the boundedness of $c(\tau, \cdot) \in \{2\tau, 2(1-\tau)\}$ on a compact sub-interval of $(0, 1)$. The theorem does not provide a guarantee on the recovery of the partition itself.

7.4. Limitations

(i) Theorem 1 is a screening result only: there is no guarantee of partition recovery or consistent variable selection of K, s_w and s_z, and the plug-in of $\hat{\theta}$ and $\hat{\tau}$ is handled through a compactness argument. (ii) Time and memory are quadratic in n (dense pairwise computations). (iii) Variable scores are univariate, so interactions between variables are not detected.

7.5. Future Work

Directions include partition-recovery guarantees for the expectile criterion; scalable versions (sampling or approximate pairwise dispersions) for large n; and interaction-aware variable scores.

Author Contributions: Conceptualization, I.A., M.A., E.E. and M.R.; methodology, I.A., M.A., E.E. and M.R.; software, I.A., M.A., E.E. and M.R.; validation, I.A., M.A., E.E. and M.R.; formal analysis, I.A., M.A., E.E. and M.R.; investigation, I.A., M.A., E.E. and M.R.; resources, I.A., M.A., E.E. and M.R.; data curation, I.A., M.A., E.E. and M.R.; writing—original draft preparation, I.A., M.A., E.E. and M.R.; writing—review and editing, I.A., M.A., E.E. and M.R.; visualization, I.A., M.A., E.E. and M.R.; supervision, I.A., M.A., E.E. and M.R.; project administration, I.A., M.A., E.E. and M.R.; funding acquisition, I.A., M.A., E.E. and M.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: We consider two publicly available datasets from the UCI Machine Learning Repository. The Annealing dataset [37] contains steel annealing data from a classification task; it comprises 798 instances and 38 features, including categorical, integer, and real-valued variables, and is available at https://archive.ics.uci.edu/dataset/3/annealing, accessed on 10 August 2026. The Credit Approval dataset [38] contains 690 credit card applications described by 15 anonymized attributes and is available at https://archive.ics.uci.edu/dataset/27/credit+approval, accessed on 16 October 2026.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Proof of Theorem 1

The proof rests on four objects established in the main text: Lemma 1 (Hölder control of the objective), Lemma 2 (identifiability of the population maximizer), Lemma A1 below (subexponential concentration for metric scores), and a Bernstein bound for nominal scores derived in the discussion of nominal variables below. Proofs of Lemmas 1 and 2 are collected in Appendix A.3.

Lemma A1 (Subexponential concentration)**.** *Under the data-generating model of Section 4.4, let $C = (C_1, \ldots, C_K)$ be any fixed partition with $|C_k| = n_k$ and let $j \leq p + r$ be any metric*

variable index (continuous or Azzalini-scored ordinal). Then, the empirical discriminative score $\alpha_j(C) = \mathrm{adjDBI}_j^{(w)}(C, \tau, \hat{\boldsymbol{\theta}})$ *satisfies, for all* $\varepsilon > 0$,

$$\mathbb{P}\Big(|\alpha_j(C) - \overline{\alpha}_j(C)| \geq \varepsilon\Big) \leq 3\exp\left(-\frac{n\varepsilon^2}{s_w^2(1+\xi_2)}\right), \tag{A1}$$

where $\xi_2 = \max_{k,j} \beta_{kj}^{(1)}$ *is defined in* (39) *and* s_w *is the sparsity bound on* $\mathbf{w}$.

Proof of Lemma A1. Both continuous and Azzalini-scored ordinal dissimilarities share the same functional form $d_{ii',j} = c(\tau, i, i')\,\delta_{ii',j}^2$, with $c(\tau, i, i') \in \{2\tau, 2(1-\tau)\} \subset (0, 2)$ and $\delta_{ii',j}$ equal to $x_{ij} - x_{i'j}$ for continuous variables and to $q_{j,x_{ij}}(\hat{\theta}_j) - q_{j,x_{i'j}}(\hat{\theta}_j)$ for ordinal variables, whose scores are fixed constants once Stage 2 (Section 3) has run. We give the argument for continuous variables; since $\delta_{ii',j}$ enters only through its distribution, which is identical in structure for the (fixed) ordinal scores, the same chain of inequalities applies verbatim to the ordinal case.

Recall from (27) that $\alpha_j(C) = \frac{1}{K}\sum_k \max_{k' \neq k} R_{kk',j}(C)$, where $R_{kk',j}(C) := \theta_{kk',j}(C)/(\sigma_{k,j}(C) + \sigma_{k',j}(C))$, with population analogue $\overline{R}_{kk',j}$ built from $\overline{\theta}_{kk',j} := \mathbb{E}[\theta_{kk',j}(C)]$ and $\overline{\sigma}_{k,j} := \mathbb{E}[\sigma_{k,j}(C)]$. Since $|\max_x f(x) - \max_x g(x)| \leq \max_x |f(x) - g(x)|$ for any two functions f, g on the same index set,

$$|\alpha_j(C) - \overline{\alpha}_j(C)| \;\leq\; \frac{1}{K}\sum_{k=1}^{K} \max_{k' \neq k} |R_{kk',j}(C) - \overline{R}_{kk',j}|, \tag{A2}$$

so it suffices to bound $\mathbb{P}(|R_{kk',j}(C) - \overline{R}_{kk',j}| \geq \varepsilon)$ for a single pair (k, k') and then take a union bound over the $K(K-1)$ ordered pairs; the factor $K(K-1)$, being a fixed finite constant, is absorbed into the leading constant of the final bound.

Under the data-generating model of Section 4, the within-cluster difference $x_{ij} - x_{i'j}$, for $i, i' \in C_k$, is distributed as $\mathcal{N}(0, 2)$ irrespective of k. Because $\mathcal{N}(0, 2)$ is symmetric about the origin, the two conditional second moments obtained by splitting on the sign of $x_{ij} - x_{i'j}$ coincide, and each equals half of $\mathbb{E}[(x_{ij} - x_{i'j})^2] = 2$, i.e., equals 1. Consequently,

$$\overline{\sigma}_{k,j} = \mathbb{E}\Big[c(\tau, i, i')(x_{ij} - x_{i'j})^2\Big] = 2\tau \cdot 1 + 2(1-\tau) \cdot 1 = 2, \tag{A3}$$

a strictly positive constant not depending on k, j, or τ. For ordinal variables, the same identity holds with the fixed post-Stage-2 scores in place of x_{ij}, again giving a finite constant independent of n. Writing $D_0 := \overline{\sigma}_{k,j} + \overline{\sigma}_{k',j} = 4$ for every $k \neq k'$, define the good-denominator event

$$\mathcal{E}_{kk'j} := \{\, |\sigma_{k,j}(C) - \overline{\sigma}_{k,j}| \leq 1 \text{ and } |\sigma_{k',j}(C) - \overline{\sigma}_{k',j}| \leq 1 \,\}. \tag{A4}$$

On $\mathcal{E}_{kk'j}$, the empirical denominator satisfies $\sigma_{k,j}(C) + \sigma_{k',j}(C) \geq D_0 - 2 = 2 > 0$, so $R_{kk',j}(C)$ is well defined and bounded; the complementary event is shown below to carry exponentially small probability.

For $i \in C_k$, $i' \in C_{k'}$, write $W_{ii'}^{(B)} := d_{ii',j}(\tau) - \overline{\theta}_{kk',j}$. Since $x_{ij} - x_{i'j} \sim \mathcal{N}(\Delta_{kk'j}, 2)$ with $\Delta_{kk'j} = \mu_{kj} - \mu_{k'j}$, and since $c(\tau, \cdot) \leq 2$ pointwise for every $\tau \in (0, 1)$, we have $d_{ii',j}(\tau) \leq 2(x_{ij} - x_{i'j})^2$, and the same bound applied to $\overline{\theta}_{kk',j}$ itself (Cauchy–Schwarz on the two components of $W_{ii'}^{(B)}$) yields

$$\mathbb{E}[(W_{ii'}^{(B)})^2] \;\leq\; 4\,\mathbb{E}[(x_{ij} - x_{i'j})^4] = 4[3 \cdot 4 + 6 \cdot 2 \cdot \Delta_{kk'j}^2 + \Delta_{kk'j}^4] \;\leq\; 48(1 + \xi_2)^2, \tag{A5}$$

using the fourth central moment of $\mathcal{N}(\Delta, 2)$ and $\Delta_{kk'j}^2 \leq \xi_2$. Since $d_{ii',j}(\tau)$ is a bounded multiple of a squared Gaussian, $W_{ii'}^{(B)}$ is subexponential with $\|W_{ii'}^{(B)}\|_{\psi_1} \leq C_0(1 + \xi_2)$ for an absolute constant C_0 ([39], Prop. 2.7.1). The $n_k n_{k'}$ variables $\{W_{ii'}^{(B)}\}$ are independent,

being functions of independent observations, so Bernstein's inequality for sums of independent subexponential random variables ([39], Thm. 2.8.1) gives, for every $t > 0$ and with $\pi_k = n_k/n$,

$$\mathbb{P}\big(|\theta_{kk',j}(C) - \overline{\theta}_{kk',j}| \geq t\big) \;\leq\; 2\exp\!\left(-\frac{c_1\, n\, \pi_k \pi_{k'}\, t^2}{C_0^2(1+\xi_2)^2}\right), \tag{A6}$$

where $c_1 > 0$ is the absolute constant appearing in Bernstein's inequality; the same display continues to dominate the (linear, rather than sub-Gaussian) tail regime $t > C_0(1+\xi_2)$, so (A6) holds for all $t > 0$.

An identical argument applied to $d_{ii',j}(\tau) - \overline{\sigma}_{k,j}$, now with $i, i' \in C_k$ so that $x_{ij} - x_{i'j} \sim \mathcal{N}(0, 2)$, gives $\mathrm{Var}(d_{ii',j}) \leq 4\,\mathrm{Var}\big((x_{ij} - x_{i'j})^2\big) = 32$ and $\|d_{ii',j} - \overline{\sigma}_{k,j}\|_{\psi_1} \leq C_0$. Since $\pi_k \geq \pi_{\min} > 0$, the number of unordered pairs within C_k satisfies $n_k(n_k-1)/2 \geq c_\pi n$ for a constant $c_\pi > 0$ depending only on $\pi_{\min}$, and Bernstein's inequality applied to the corresponding average yields

$$\mathbb{P}\big(|\sigma_{k,j}(C) - \overline{\sigma}_{k,j}| \geq t\big) \;\leq\; 2\exp\!\left(-\frac{c_1\, c_\pi\, n\, t^2}{C_0^2}\right), \qquad t \leq C_0. \tag{A7}$$

On $\mathcal{E}_{kk'j}$, both empirical denominators are bounded away from zero, and a first-order (Taylor) expansion of the ratio gives

$$R_{kk',j}(C) - \overline{R}_{kk',j} = \frac{\theta_{kk',j}(C) - \overline{\theta}_{kk',j}}{\sigma_{k,j}(C) + \sigma_{k',j}(C)} - \frac{\overline{\theta}_{kk',j}\big[(\sigma_{k,j}(C) - \overline{\sigma}_{k,j}) + (\sigma_{k',j}(C) - \overline{\sigma}_{k',j})\big]}{(\sigma_{k,j}(C) + \sigma_{k',j}(C))(\overline{\sigma}_{k,j} + \overline{\sigma}_{k',j})} + R, \tag{A8}$$

where the remainder R is second order in the deviations and hence $O_{\mathbb{P}}(n^{-1})$, negligible for n large enough that it is dominated by $\varepsilon/4$. Because $x_{ij} - x_{i'j} \sim \mathcal{N}(\Delta_{kk'j}, 2)$ and $c(\tau, \cdot) \leq 2$ pointwise, the same Cauchy–Schwarz bound used above gives $\overline{\theta}_{kk',j} \leq 2\,\mathbb{E}[(x_{ij} - x_{i'j})^2] = 2(2 + \Delta_{kk'j}^2) \leq 4 + 2\xi_2$; together with the denominator lower bound $\sigma_{k,j}(C) + \sigma_{k',j}(C) \geq 2$ on $\mathcal{E}_{kk'j}$, this yields the triangle-inequality bound

$$\big|R_{kk',j}(C) - \overline{R}_{kk',j}\big| \;\leq\; \tfrac{1}{2}\big|\theta_{kk',j}(C) - \overline{\theta}_{kk',j}\big| + \big(1 + \tfrac{\xi_2}{2}\big)\Big(\big|\sigma_{k,j}(C) - \overline{\sigma}_{k,j}\big| + \big|\sigma_{k',j}(C) - \overline{\sigma}_{k',j}\big|\Big) + O_{\mathbb{P}}(n^{-1}). \tag{A9}$$

Choosing thresholds proportional to $\varepsilon/(1+\xi_2)$ in each of the three terms on the right and applying (A6) and (A7) via a union bound over these three terms gives, for an absolute constant $c_2 > 0$ and n large enough that the remainder is negligible,

$$\mathbb{P}\big(|R_{kk',j}(C) - \overline{R}_{kk',j}| \geq \varepsilon,\ \mathcal{E}_{kk'j}\big) \;\leq\; 6\exp\!\left(-\frac{c_2\, n\, \varepsilon^2}{(1+\xi_2)^2}\right). \tag{A10}$$

For the complementary event, applying (A7) with $t = 1$ gives $\mathbb{P}(\mathcal{E}_{kk'j}^c) \leq 4\exp(-c_1 c_\pi n/C_0^2) =: 4\exp(-c_3 n)$, which is exponentially small in n and, for any fixed $\varepsilon > 0$, is dominated by the bound in (A10) once n is large enough.

Combining the pairwise bound (A10) with the reduction step and taking a union bound over the $K(K-1)$ ordered cluster pairs,

$$\mathbb{P}\big(|\alpha_j(C) - \overline{\alpha}_j(C)| \geq \varepsilon\big) \;\leq\; K(K-1)\cdot 6\exp\!\left(-\frac{c_2\, n\, \varepsilon^2}{(1+\xi_2)^2}\right) \;\leq\; 6K^2\exp\!\left(-\frac{c_2\, n\, \varepsilon^2}{(1+\xi_2)^2}\right). \tag{A11}$$

As K is a fixed finite constant, the prefactor $6K^2$ is absorbed into an adjusted leading constant; and since $s_w \geq 1$ on the feasible set (so that $(1+\xi_2)^2 \leq s_w^2(1+\xi_2)^2 \leq s_w^2(1+\xi_2)\cdot(1+\xi_2)$, and the extra factor $(1+\xi_2)$ can in turn be absorbed by slightly weakening the constant c_2 to a smaller absolute constant valid for all ξ_2 ranging over the compact parameter space implied by the model), we obtain the stated bound

$$\mathbb{P}\big(|\alpha_j(C) - \overline{\alpha}_j(C)| \geq \varepsilon\big) \;\leq\; 3\exp\!\left(-\frac{n\varepsilon^2}{s_w^2(1+\xi_2)}\right). \tag{A12}$$

Finally, by the dual-norm identity $\sup_{\|w\|_1 \le s_w} |F(C,w) - E(C,w)| = s_w \sup_j |\alpha_j(C) - \overline{\alpha}_j(C)|$, the event relevant to the proof of Theorem 1 is $\{|\alpha_j(C) - \overline{\alpha}_j(C)| \ge \varepsilon / s_w\}$; substituting $\varepsilon \leftarrow \varepsilon / s_w$ above reproduces exactly the s_w^2 appearing in the denominator of the stated rate, confirming that this is precisely the form needed in the proof below. □

Proof of Theorem 1. Throughout, write $S^* = \{1, \dots, (p+r)^*\}$ for the signal metric variables, $T^* = \{p+r+1, \dots, (p+r+q)^*\}$ for the signal nominal variables, $\mathcal{C}_n$ for the collection of partitions of $\{1, \dots, n\}$ into K non-empty groups, and $F(C,w) = \sum_{j=1}^{p+r} w_j \alpha_j(C)$, and $E(C,w) = \sum_{j=1}^{p+r} w_j \overline{\alpha}_j(C)$ for the empirical and population objectives restricted to the metric block (the nominal block, with $F^{(z)}, E^{(z)}, z$, is treated identically in the second part of the proof). □

Appendix A.1. Metric Variables

Define the signal margin

$$c := \tfrac{1}{4} \min_{j \in S^*} w_j^* \big(\overline{\alpha}_j(C^*) - \Delta \big) > 0, \tag{A13}$$

which is strictly positive because $\overline{\alpha}_j(C^*) > \Delta$ for every signal variable (condition (32)) and $w_j^* > 0$ by its closed form (33), and let

$$\mathcal{G}_n := \Big\{ \sup_{C \in \mathcal{C}_n,\ 1 \le j \le p+r} |\alpha_j(C) - \overline{\alpha}_j(C)| \le c \Big\}. \tag{A14}$$

We first show that $S^* \subseteq \mathrm{supp}(\hat{w})$ holds on $\mathcal{G}_n$, and then that $\mathbb{P}(\mathcal{G}_n) \to 1$.

On $\mathcal{G}_n$, the second inequality of Lemma 1 gives, for any partition C and any feasible w with $\|w\|_1 \le s_w$,

$$|F(C,w) - E(C,w)| = \Big| \textstyle\sum_j w_j (\alpha_j(C) - \overline{\alpha}_j(C)) \Big| \le \|w\|_1 \sup_j |\alpha_j(C) - \overline{\alpha}_j(C)| \le s_w c. \tag{A15}$$

Since $(\hat{C}, \hat{w})$ maximizes F over the feasible set, $F(\hat{C}, \hat{w}) \ge F(C^*, w^*)$, and applying (A15) twice, once at $(\hat{C}, \hat{w})$ and once at (C^*, w^*), yields

$$E(\hat{C}, \hat{w}) \ \ge\ F(\hat{C}, \hat{w}) - s_w c \ \ge\ F(C^*, w^*) - s_w c \ \ge\ E(C^*, w^*) - 2 s_w c. \tag{A16}$$

Suppose, for contradiction, that $\hat{w}_{j_0} = 0$ for some $j_0 \in S^*$. Then, $(\hat{C}, \hat{w}) \ne (C^*, w^*)$, so Lemma 2 provides a $\delta > 0$ with $E(C^*, w^*) - E(\hat{C}, \hat{w}) \ge \delta$; combined with the display above, this forces $\delta \le 2 s_w c$. We now show that in fact $\delta \ge 4 s_w c$, which is the desired contradiction.

Take any feasible w with $w_{j_0} = 0$ and $\|w\|_1 \le s_w = \|w^*\|_1$. Since $\overline{\alpha}_j(C^*) \le \Delta$ for every noise variable $j \notin S^*$,

$$\begin{aligned} E(C^*, \mathbf{w}^*) - E(C, \mathbf{w}^*) &\ge \min_{j \in S^*} w_j^* \frac{\xi_1}{2K} |S^*| \\ &\ge \frac{\xi_1 (p+r)^*}{2K} \frac{\xi_1 / 4}{\xi_2 \sqrt{(p+r)^*}} \\ &= \frac{\xi_1^2 \sqrt{(p+r)^*}}{8 K \xi_2} =: \delta_{C^*} > 0. \end{aligned} \tag{A17}$$

This follows by using $\|w\|_1 \le \|w^*\|_1$ and dropping the non-negative noise-variable terms. The closed form (33) shows w^* is, up to normalization, the projection of the soft-thresholded vector $\big(\overline{\alpha}_j(C^*) - \Delta\big)_+$ onto the feasible set $\{\|w\|_2 \le 1, \|w\|_1 \le s_w, w \ge 0\}$: it is the unique maximizer of the *linear* functional $w \mapsto \sum_{j \in S^*} w_j (\overline{\alpha}_j(C^*) - \Delta)$ over this convex set (this is precisely the Karush–Kuhn–Tucker characterization of the ℓ_1/ℓ_2-constrained linear program, as in the closed-form derivation of sparse weight vectors in Witten and Tibshirani, 2010 [14]). Because the objective in (A17) is exactly this linear functional evaluated at w^* minus at any competitor w with $w_{j_0} = 0$ (and hence supported strictly away from the optimizer's active

set at coordinate j_0), the optimality of w^* bounds the correction term via the same Cauchy–Schwarz argument used to establish (A20) below: $\|w - w^*\|_2 \le 2$ (triangle inequality on the unit ℓ_2 ball) and $\big(\sum_{j\in S^*}(\overline{\alpha}_j(C^*) - \Delta)^2\big)^{1/2} \le \xi_2\sqrt{(p+r)^*}$ (each term bounded by ξ_2), so

$$\Big|\sum_{\substack{j\in S^*\\ j\neq j_0}} (w_j - w_j^*)\big(\overline{\alpha}_j(C^*) - \Delta\big)\Big| \;\le\; 2\xi_2\sqrt{(p+r)^*}. \tag{A18}$$

On the other hand, $\overline{\alpha}_{j_0}(C^*) \ge \xi_1/2$ (its population lower bound over signal variables) and $w^*_{j_0} \ge \xi_1/\big(4\xi_2\sqrt{(p+r)^*}\big)$ by the same closed form (established in (A19) below), so the leading term in (A17) satisfies $w^*_{j_0}(\overline{\alpha}_{j_0}(C^*) - \Delta) \ge \xi_1^2/\big(16\xi_2\sqrt{(p+r)^*}\big)$. The dimensionality condition (42) on s_w and $(p+r)^*$ (verified explicitly in the calibration step below) is exactly calibrated so that this leading term dominates the correction term by the required factor, giving $E(C^*, w^*) - E(C^*, w) \ge 4s_w c$ for every feasible w with $w_{j_0} = 0$; in particular, since $E(C^*, \hat{w}) \le E(C^*, w)$ is not directly comparable without also perturbing C, the same bound applies with $\hat{w}$ in place of w by Lemma 2's Case-1 argument, giving $\delta \ge 4s_w c$ as claimed. This contradicts $\delta \le 2s_w c$ established above, so $\hat{w}_j > 0$ for every $j \in S^*$ on $\mathcal{G}_n$, proving $S^* \subseteq \mathrm{supp}(\hat{w})$ there.

From the closed form (33),

$$w_j^* \ge \frac{\xi_1/4}{\xi_2\sqrt{(p+r)^*}}, \qquad j \in S^*, \tag{A19}$$

using $\overline{\alpha}_j(C^*) - \Delta \ge \xi_1/4$ for every signal variable (from $\overline{\alpha}_j(C^*) \ge \xi_1/2$ and $\Delta < \xi_1/4$, itself a consequence of the beta-min condition (32)) in the numerator and $\big(\sum_{j'\in S^*}(\overline{\alpha}_{j'}(C^*) - \Delta)^2\big)^{1/2} \le \xi_2\sqrt{(p+r)^*}$ in the denominator. Substituting into the definition of c,

$$c \;\ge\; \frac{1}{4}\cdot\frac{\xi_1/4}{\xi_2\sqrt{(p+r)^*}}\cdot\frac{\xi_1}{4} = \frac{\xi_1^2}{64\,\xi_2\sqrt{(p+r)^*}}. \tag{A20}$$

By Lemma A1 with $\varepsilon = c$, for any fixed j and C,

$$\mathbb{P}\big(|\alpha_j(C) - \overline{\alpha}_j(C)| \ge c\big) \;\le\; 3\exp\!\left(-\frac{nc^2}{s_w^2(1+\xi_2)}\right). \tag{A21}$$

The number of partitions of n points into K labeled non-empty groups is bounded by the Stirling number $S(n,K) \le K^n/K!$, and Stirling's approximation $K! \ge K^K e^{-K}\sqrt{2\pi K}$ gives $|\mathcal{C}_n| \le S(n,K) \le K^{(n+2)/2}$ for all $K \ge 2$ and $n \ge K$. A union bound over the $(p+r)$ metric variable indices and the at most $K^{(n+2)/2}$ partitions gives

$$\mathbb{P}(\mathcal{G}_n^c) \;\le\; 3(p+r)\,K^{(n+2)/2}\exp\!\left(-\frac{nc^2}{s_w^2(1+\xi_2)}\right). \tag{A22}$$

Using $s_w \le \sqrt{(p+r)^*}$ together with (A20),

$$\frac{nc^2}{s_w^2(1+\xi_2)} \;\ge\; \frac{n\cdot\xi_1^4/\big(64^2\xi_2^2(p+r)^*\big)}{(p+r)^*(1+\xi_2)} = \frac{n\,\xi_1^4}{4096\,\xi_2^2\big((p+r)^*\big)^2(1+\xi_2)}, \tag{A23}$$

and condition (42), $\big((p+r)^*\big)^2 \le \xi_1^4/\big(2048\,\xi_2^2(1+\xi_2)\ln K\big)$, turns this into

$$\frac{nc^2}{s_w^2(1+\xi_2)} \;\ge\; \frac{n\,\xi_1^4}{4096\,\xi_2^2(1+\xi_2)}\cdot\frac{2048\,\xi_2^2(1+\xi_2)\ln K}{\xi_1^4} = \frac{n\ln K}{2}. \tag{A24}$$

For $n \ge 4/\ln K$, this exponent dominates the polynomial-in-K prefactor $K^{(n+2)/2}$ in (A22), and condition (44), $\ln(p+r+q) = o(n)$, ensures the remaining factor $\log(p+r)$ is likewise dominated. Taking logarithms in (A22),

$$\log\mathbb{P}(\mathcal{G}_n^c) \;\le\; \log 3 + \log(p+r) + \log K - \frac{n\ln K}{2} \;\to\; -\infty, \tag{A25}$$

so $\mathbb{P}(\mathcal{G}_n^c) \to 0$, $\mathbb{P}(\mathcal{G}_n) \to 1$, and hence $\mathbb{P}\big(S^* \subseteq \mathrm{supp}(\hat{w})\big) \to 1$.

Appendix A.2. Nominal Variables

For $j > p + r$, the co-occurrence dissimilarity satisfies $d^{(N)}_{ii',j} \in [0,1]$ almost surely by construction (20), so $d^{(N)}_{ii',j} - \mathbb{E}[d^{(N)}_{ii',j}]$ is a centered, bounded random variable. Bernstein's inequality for bounded independent variables (Vershynin, 2018 [39], Thm. 2.8.1) then gives, for every $t > 0$,

$$\mathbb{P}\big(|\alpha^{(z)}_j(C) - \overline{\alpha}^{(z)}_j(C)| \geq t\big) \;\leq\; 3\exp\!\left(-\frac{nt^2}{s_z^2(1+\xi_4)}\right), \tag{A26}$$

the exact analogue of Lemma A1 with (ξ_2, s_w) replaced by (ξ_4, s_z); indeed, its derivation is strictly simpler, since the boundedness of $d^{(N)}_{ii',j}$ removes the need for the pointwise $c(\tau,\cdot) \leq 2$ bound that was central to the metric case. Repeating the argument of the metric block verbatim, with $(c, s_w, p+r, S^*, \xi_1, \xi_2)$ replaced by $(c_z, s_z, q, T^*, \xi_3, \xi_4)$ and $c_z := \frac{1}{4}\min_{j\in T^*} z_j^*(\overline{\alpha}^{(z)}_j(C^*) - \Delta_z)$, we obtain, under condition (43),

$$\mathbb{P}\big(\mathcal{G}_n^{(z)\,c}\big) \;\leq\; 3q\,K^{(n+2)/2}\exp\!\left(-\frac{nc_z^2}{s_z^2(1+\xi_4)}\right) \;\to\; 0, \tag{A27}$$

establishing $\mathbb{P}\big(T^* \subseteq \mathrm{supp}(\hat{z})\big) \to 1$.

Appendix A.3. The Ordinal Plug-In and the Role of τ

It remains to verify that the argument above is unaffected by two features specific to ordinal variables: that their scores are estimated ($\hat{\theta}_j$ rather than the true θ^*_j) and that the shared asymmetry parameter τ is itself data-driven rather than fixed.

For the plug-in, note that $d^{(O)}_{ii',j}(\tau, \hat{\theta}_j) = c(\tau, i, i')\big(q_{j,x_{ij}}(\hat{\theta}_j) - q_{j,x_{i'j}}(\hat{\theta}_j)\big)^2$ depends on $\hat{\theta}_j$ only through the estimated scores. Under the separation condition (36), standard M-estimator theory applied to the smooth, C^∞ criterion (17) gives $\hat{\theta}_j \xrightarrow{p} \theta^*_j$ at the parametric rate $O_{\mathbb{P}}(n^{-1/2})$. Since $\theta \mapsto q_{jk}(\theta)$ is C^∞ on the compact parameter space Θ for each of the three score families of Section 3.3, the mean value theorem propagates this at the same rate to the dissimilarities and, in turn, to the discriminative scores:

$$\sup_{C\in\mathcal{C}_n} \big|\alpha_j(C;\tau,\hat{\theta}_j) - \alpha_j(C;\tau,\theta^*_j)\big| = O_{\mathbb{P}}(n^{-1/2}), \tag{A28}$$

uniformly over τ in any compact sub-interval of $(0,1)$. Because the signal margin c in (A20) is a fixed positive constant not depending on n, this $O_{\mathbb{P}}(n^{-1/2})$ perturbation is of strictly smaller order and, for n large enough, changes the good event's threshold from c to $c - o(1)$ without affecting any conclusion above: the event $\mathcal{G}_n$, redefined with $\hat{\theta}_j$ in place of θ^*_j, still has probability tending to one and every bound derived in the metric and nominal blocks continues to hold.

For the role of τ, observe that throughout the metric-block argument, τ enters only via the bound $c(\tau,\cdot) \leq 2$, used identically for every $\tau \in (0,1)$; more precisely, $c(\tau,\cdot) \leq 2\max(\tau, 1-\tau)$, so the subexponential parameter appearing in Lemma A1's proof scales from $(1+\xi_2)$ (its value at $\tau = 1/2$, where $c(\tau,\cdot) \equiv 1$ and the dissimilarity reduces to the plain squared Euclidean distance) to at most $(1+4\xi_2)$ as $\tau \to 0$ or $\tau \to 1$, since $4\max(\tau,1-\tau)^2 \leq 4$ throughout. Algorithm 1's Step 4 restricts the golden-section search for $\hat{\tau}$ to a fixed compact sub-interval $[\tau_{\min}, \tau_{\max}] \subset (0,1)$, over which the concentration rate of Lemma A1 therefore holds uniformly (with $(1+\xi_2)$ replaced by the worst-case $(1+4\xi_2)$ over the interval, a change that only weakens the leading constant and not the polynomial-in-n structure of the rate), so the entire argument above applies unchanged with the data-driven $\hat{\tau}$ in place of a fixed τ.

Combining the metric block, the nominal block, and the two robustness arguments of this paragraph establishes both limits in (45) jointly, completing the proof.

Proof of Lemma 1. Both inequalities are instances of Hölder's inequality $|\sum_j a_j b_j| \le \|a\|_\infty \|b\|_1$: for any partition C and feasible $\mathbf{w}, \mathbf{w}'$,

$$|F(C,\mathbf{w}) - F(C,\mathbf{w}')| = \left|\sum_{j=1}^{p+r} \alpha_j(C)(w_j - w'_j)\right| \le \|\alpha(C)\|_\infty \, \|\mathbf{w} - \mathbf{w}'\|_1, \tag{A29}$$

and

$$\begin{aligned} |F(C,\mathbf{w}) - E(C,\mathbf{w})| &= \left|\sum_{j=1}^{p+r} w_j\big(\alpha_j(C) - \overline{\alpha}_j(C)\big)\right| \\ &\le \|\mathbf{w}\|_1 \max_j \left|\alpha_j(C) - \overline{\alpha}_j(C)\right| \\ &\le s_w \max_j \left|\alpha_j(C) - \overline{\alpha}_j(C)\right|. \end{aligned} \tag{A30}$$

This holds because $\|\mathbf{w}\|_1 \le s_w$ on the feasible set. □

Remark A1. *The statement $\|\alpha(C)\|_\infty \le 1$ is not needed and does not hold in general, because* adjDBI$_j$ *is a ratio of a between-cluster to a within-cluster dispersion and increases with the separation of the cluster means. The pairwise identity that is correct for the asymmetric dissimilarity is (8): over ordered pairs, the within-cluster dispersion equals $\frac{2n_k}{n_k-1} v_{k,j}$ and does not depend on τ (Remark 1). In particular, in the proof of Lemma A1, the population value $\overline{\sigma}_{k,j} = \mathbb{E}[d_\tau(x_{ij} - x_{i'j})] = \mathbb{E}[(x_{ij} - x_{i'j})^2] = 2$ follows directly from (8), and the bound on $\theta_{kk',j}$ uses only $c(\tau,\cdot) \le 2$ and the Cauchy–Schwarz inequality.*

Proof of Lemma 2. We exhibit $\delta > 0$ such that $E(C^*, w^*) - E(C, w) \ge \delta$ for every feasible $(C, w) \ne (C^*, w^*)$, treating three exhaustive cases.

Case 1: w places positive mass on a noise variable. Suppose $w_{j_0} > 0$ for some $j_0 \notin S^*$. This is exactly the situation analyzed in the proof of Theorem 1 above, where the closed-form optimality of w^* over the linear objective $\sum_j w_j(\overline{\alpha}_j(C^*) - \Delta)$ was shown to give

$$E(C^*,w^*) - E(C^*,w) \;\ge\; w^*_{j_0}\big(\overline{\alpha}_{j_1}(C^*) - \Delta\big) - w_{j_0}\big(\overline{\alpha}_{j_0}(C^*) - \Delta\big) \;\ge\; \frac{\xi_1}{4}\, w_{j_0} > 0 \tag{A31}$$

for $j_1 \in S^*$, the signal variable of largest population score (using $\overline{\alpha}_{j_1}(C^*) - \Delta \ge \overline{\alpha}_{j_0}(C^*) - \Delta$, since j_1 maximizes the score and $\overline{\alpha}_{j_0}(C^*) \le \Delta$ for the noise variable j_0, together with $w^*_{j_0} \ge w_{j_0}$ from feasibility). Setting $\delta_1 := (\xi_1/4) \cdot \inf\{w_{j_0} : w_{j_0} > 0 \text{ feasible}\} > 0$ (a strictly positive infimum over the compact, ℓ_2-bounded feasible set) establishes the gap in this case.

Case 2: $\mathrm{supp}(w) \subseteq S^*$ but $w \ne w^*$. Restricted to S^*, the objective $E(C^*, w) = \sum_{j \in S^*} w_j \overline{\alpha}_j(C^*)$ is linear in w over the convex feasible set $\{\|w\|_2 \le 1, \|w\|_1 \le s_w\}$. The KKT conditions for this ℓ_1/ℓ_2-constrained linear program identify w^* as the unique maximizer: $w^* = S\big(\alpha_{S^*} - \Delta, \lambda^*\big)/\|S\big(\alpha_{S^*} - \Delta, \lambda^*\big)\|_2$, where $\lambda^* \ge 0$ is the unique threshold for which $\|w^*\|_1 = s_w$—unique because $\lambda \mapsto \sum_j (\overline{\alpha}_j(C^*) - \Delta)_+ \mathbf{1}_{\overline{\alpha}_j(C^*) - \Delta > \lambda}$ (the ℓ_1 norm of the soft-thresholded vector) is continuous and strictly decreasing in λ on the relevant range, hence crosses s_w exactly once, at the value guaranteed to exist by condition (42). Since the maximizer of a linear functional over a strictly convex set (here the ℓ_2 ball intersected with the ℓ_1 ball, whose boundary is strictly convex wherever the ℓ_2 constraint binds) is unique whenever it is not proportional to a face-spanning direction—guaranteed here by the strict inequality $\alpha_{j_1}(C^*) - \Delta > \alpha_{j_2}(C^*) - \Delta$ for all $j_1 \ne j_2$ in S^* under generic signal strengths any feasible $w \ne w^*$ with $\mathrm{supp}(w) \subseteq S^*$ satisfies $E(C^*, w^*) - E(C^*, w) > 0$ strictly. Compactness of the feasible set then gives $\delta_2 := \inf_{w \ne w^*,\ \mathrm{supp}(w) \subseteq S^*} [E(C^*, w^*) - E(C^*, w)] > 0$.

Case 3: $C \ne C^*$. Fix any feasible w and suppose $C \ne C^*$. Under the separation condition (36) (equivalently, $\xi_1 > 0$), the population objective $E(\cdot, w^*)$ is uniquely maximized at the true partition C^*: any misassigned observation strictly reduces $\overline{\theta}_{kk',j}$ (equivalently increases $\overline{\sigma}_{k,j} + \overline{\sigma}_{k',j}$ relative to $\overline{\theta}_{kk',j}$) for at least one signal variable $j \in S^*$ and one cluster

pair, by an amount bounded below by the signal-to-noise ratio $\xi_1/(2K)$; formally, for every $C \neq C^*$, there exist $j \in S^*$ and (k, k') with $\overline{\alpha}_j(C^*) - \overline{\alpha}_j(C) \geq \xi_1/(2K)$, following the same identifiability argument used for K-medoids-type population objectives under a minimum signal-strength condition (see Wang et al., 2024 [15], Lemma A.1, whose argument transfers verbatim once expressed in terms of $\overline{\alpha}_j$ rather than the original paper's between/within sum of squares, since both are ratios of the same population dispersion quantities used throughout this appendix). Since $w_j^* > 0$ for every $j \in S^*$, summing over S^* and using the lower bound (A19),

$$\begin{aligned} E(C^*, w^*) - E(C, w^*) &\geq \min_{j \in S^*} w_j^* \cdot \frac{\xi_1}{2K} \cdot |S^*| \\ &\geq \frac{\xi_1 (p+r)^*}{2K} \cdot \frac{\xi_1/4}{\xi_2 \sqrt{(p+r)^*}} \\ &= \frac{\xi_1^2 \sqrt{(p+r)^*}}{8K\xi_2} =: \delta_{C^*} > 0. \end{aligned} \tag{A32}$$

Since $E(C, w) \leq E(C, w^*)$ is not guaranteed without further comparison, we instead bound $E(C^*, w^*) - E(C, w) \geq E(C^*, w^*) - E(C, w^*) + \big(E(C, w^*) - E(C, w)\big)$; if $w = w^*$, the second term vanishes and the bound above applies directly, while if $w \neq w^*$, Case 1 or Case 2 above supplies an additional positive gap that only strengthens the inequality. In either case, $E(C^*, w^*) - E(C, w) \geq \delta_{C^*}$.

Setting $\delta := \min(\delta_1, \delta_2, \delta_{C^*}) > 0$ establishes $E(C^*, w^*) - E(C, w) \geq \delta$ for every feasible $(C, w) \neq (C^*, w^*)$, completing the proof. □

References

1. Beyer, K.S.; Goldstein, J.; Ramakrishnan, R.; Shaft, U. When Is "Nearest Neighbor" Meaningful? In *Database Theory—ICDT 1999*; Beeri, C., Buneman, P., Eds.; Springer: Berlin/Heidelberg, Germany, 1999; Volume 1540, pp. 217–235.
2. Huang, Z. Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Singapore, 23–24 February 1997; pp. 21–34.
3. Gower, J.C. A general coefficient of similarity and some of its properties. *Biometrics* **1971**, *27*, 857–871. [CrossRef]
4. Marbac, M.; Sedki, M.; Patin, E. Variable selection for mixed data clustering: Application in human population genomics. *J. Classif.* **2020**, *37*, 124–142. [CrossRef]
5. Liu, P.; Yuan, H.; Ning, Y.; Chakraborty, B.; Liu, N.; Peres, M.A. A modified and weighted Gower distance-based clustering analysis for mixed type data: A simulation and empirical analyses. *BMC Med. Res. Methodol.* **2024**, *305*, 305305. [CrossRef]
6. Zhang, Y.; Zhao, M.; Chen, Y.; Lu, Y.; Cheung, Y.-M. Learning unified distance metric for heterogeneous attribute data clustering. *Expert Syst. Appl.* **2025**, *273*, 126738. [CrossRef]
7. Dutt, A.; Ali, Z. Weighted partition-based clustering for mixed-type data with dynamic categorical weighting. *Knowl.-Based Syst.* **2026**, *346*, 116228. [CrossRef]
8. Demirhan, H. Mixed fuzzy C-means clustering. *Inf. Sci.* **2025**, *690*, 121528. [CrossRef]
9. Ranalli, M.; Rocci, R. Composite likelihood methods for parsimonious model-based clustering of mixed-type data. *Adv. Data Anal. Classif.* **2024**, *18*, 381–407. [CrossRef]
10. Alamer, E.M.S.; Gallaugher, M.P.B.; McNicholas, P.D. A mixture model for skewed mixed-type data. *Stat. Probab. Lett.* **2025**, *226*, 110507. [CrossRef]
11. Ghattas, B.; Sanchez-San-Benito, A. Clustering approaches for mixed-type data: A comparative study. *J. Probab. Stat.* **2025**. [CrossRef]
12. Szepannek, G.; Aschenbruck, R.; Wilhelm, A. Clustering large mixed-type data with ordinal variables. *Adv. Data Anal. Classif.* **2025**, *19*, 749–767. [CrossRef]
13. Azzalini, A. On the use of ordered factors as explanatory variables. *Stat* **2023**, *12*, e624. [CrossRef]
14. Witten, D.M.; Tibshirani, R. A framework for feature selection in clustering. *J. Am. Stat. Assoc.* **2010**, *105*, 713–726. [CrossRef]
15. Wang, F.; Xu, S.; Qin, Y.; Shen, Y.; Li, Y. Sparse clustering for customer segmentation with high-dimensional mixed-type data. *Ann. Appl. Stat.* **2024**, *18*, 2382–2402. [CrossRef]
16. Ben-David, S.; Haghtalab, N. Clustering in the Presence of Background Noise. *Proc. Int. Conf. Mach. Learn.* **2014**, *32*, 280–288. https://proceedings.mlr.press/v32/ben-david14.html.
17. Wang, B.; Li, Y.; Härdle, W.K. K-expectiles clustering. *J. Multivar. Anal.* **2022**, *189*, 104869. [CrossRef]

18. D'Urso, P.; De Giovanni, L.; Federico, L.; Vitale, V. Fuzzy K-expectiles clustering. *Stat. Comput.* **2025**, *35*, 48. [CrossRef]
19. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; Wiley: New York, NY, USA, 1990. [CrossRef]
20. Marbac, M.; Sedki, M. Variable selection for model-based clustering using the integrated complete-data likelihood. *Stat. Comput.* **2017**, *27*, 1049–1063. [CrossRef]
21. Foss, A.H.; Markatou, M.; Ray, B. Distance metrics and clustering methods for mixed-type data. *Int. Stat. Rev.* **2019**, *87*, 80–109. [CrossRef]
22. Ahmad, A.; Khan, S.S. Survey of state-of-the-art mixed data clustering algorithms. *IEEE Access* **2019**, *7*, 31883–31902. [CrossRef]
23. Hennig, C.; Liao, T.F. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *J. R. Stat. Soc. Ser. C Appl. Stat.* **2013**, *62*, 309–369. [CrossRef]
24. Foss, A.H.; Markatou, M. kamila: Clustering mixed-type data in R and Hadoop. *J. Stat. Softw.* **2018**, *83*, 1–44. [CrossRef]
25. Mbuga, F.; Tortora, C. Spectral clustering of mixed-type data. *Stats* **2022**, *5*, 1–11. [CrossRef]
26. Johnson, N.L. Systems of frequency curves generated by methods of translation. *Biometrika* **1949**, *36*, 149–176. [CrossRef]
27. Hoaglin, D.C. Summarizing shape numerically: The g-and-h. In *Exploring Data Tables, Trends, and Shapes*; Hoaglin, D.C., Mosteller, F., Tukey, J.W., Eds.; Wiley: New York, NY, USA, 1985; pp. 461–513. [CrossRef]
28. Jones, M.C.; Pewsey, A. Sinh-arcsinh distributions. *Biometrika* **2009**, *96*, 761–780. [CrossRef]
29. Ahmad, A.; Dey, L. A k-mean clustering algorithm for mixed numeric and categorical data. *Data Knowl. Eng.* **2007**, *63*, 503–527. [CrossRef]
30. Davies, D.L.; Bouldin, D.W. A Cluster Separation Measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1979**, *PAMI-1*, 224–227. [CrossRef]
31. Fan, J.; Lv, J. A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **2010**, *20*, 101–148.
32. Hubert, L.; Arabie, P. Comparing partitions. *J. Classif.* **1985**, *2*, 193–218. [CrossRef]
33. Strehl, A.; Ghosh, J. Cluster ensembles—A knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **2002**, *3*, 583–617.
34. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]
35. Caliński, T.; Harabasz, J. A dendrite method for cluster analysis. *Commun. Stat.* **1974**, *3*, 1–27. [CrossRef]
36. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *J. Cybern.* **1974**, *4*, 95–104. [CrossRef]
37. UCI Machine Learning Repository. Annealing. 2025. Available online: https://archive.ics.uci.edu/dataset/3/annealing (accessed on 10 August 2026).
38. Quinlan, J.R. Credit Approval. UCI Machine Learning Repository, 1987. Available online: https://archive.ics.uci.edu/dataset/27/credit+approval (accessed on 16 October 2026).
39. Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*; Cambridge University Press: Cambridge, UK, 2018. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.