Pattern Recognition 149 (2024) 110231

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

TriSig: Evaluating the statistical significance of triclusters

Leonardo Alexandre [a,b,c], Rafael S. Costa [c], Rui Henriques [a,b,*]

[a] INESC-ID, R. Alves Redol 9, 1000-029, Lisboa, Portugal
[b] Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal
[c] LAQV-REQUIMTE, Department of Chemistry, NOVA School of Science and Technology, Universidade NOVA de Lisboa, Campus Caparica, 2829-516, Caparica, Portugal

ARTICLE INFO

Dataset link: https://github.com/JupitersMight/TriSig

Keywords:
Triclustering
Pattern discovery
Statistical significance
Temporal pattern mining
Multivariate time series data

ABSTRACT

Tensor data analysis allows researchers to uncover novel patterns and relationships that cannot be obtained from tabular data alone. The information inferred from multi-way patterns can offer valuable insights into disease progression, bioproduction processes, behavioral responses, weather fluctuations, or social dynamics. However, spurious patterns often hamper this process. This work aims at proposing a statistical frame to assess the probability of patterns in tensor data to deviate from null expectations, extending well-established principles for assessing the statistical significance of patterns in tabular data. A principled discussion on binomial testing to mitigate false positive discoveries is entailed at the light of: variable dependencies, temporal associations and misalignments, and multi-hypothesis correction. Results gathered from the application of triclustering algorithms over distinct real-world case studies in biotechnological domains confer validity to the proposed statistical frame while revealing vulnerabilities of reference triclustering searches. The proposed assessment can be incorporated into existing triclustering algorithms to minimize spurious occurrences, rank patterns, and further prune the search space, reducing their computational complexity.

1. Introduction

In recent years, the increased capacity for comprehensively monitoring systems' behavior led to the emergence of tensor data structures such as three-dimensional data, also known as cubic or three-way data [1]. The nature of tensor data allows for a more comprehensive understanding of the underlying processes and relationships, proving to be a valuable source for knowledge acquisition and decision support [2,3]. Real-world examples in the biological domain include omic data to study how treatments affect a given tissue [4], improve plantations' health against drought [5], understand development [6], amongst other ends [7]. In social studies, user behavior has been monitored to understand group dynamics such as user preferences [8] and interactions [9]. Geophysical data has been considered to study weather patterns [10], seismic activity [11], or precision agriculture [12]. Finally, in the medical field, pattern discovery from health records has been considered for personalized medicine [13], understanding diseases [14], and neuroscience advances [15].

Although many approaches for pattern discovery in tensor data are available [16], we focus on triclustering [1], a well-established task for extracting actionable patterns within tensor data [17]. Triclustering belongs to the same branch of unsupervised tasks as clustering and biclustering. But contrary to clustering, where clusters represent groups of observations meaningfully correlated along the overall feature space, and biclustering, only prepared to handle tabular data, triclustering can make full use of the tensor data properties. Patterns extracted by triclustering algorithms, referred to as triclusters, are defined by a set of observations meaningfully correlated on a subset of variables along a subset of contexts. Table 1 synthesizes some application domains with notable ongoing contributions driven by triclustering. Triclusters are generally required to satisfy specific: *homogeneity* criteria to ensure their informative value; *dissimilarity* criteria to prevent redundancies; and *discriminative* criteria in supervised settings to leverage their predictive power [3]. However, such criteria are insufficient to guarantee the statistical significance and subsequent actionability of triclusters. To prevent spurious discoveries, the probability of observing a given tricluster when considering null expectations should be unexpectedly low. Understandably, pursuing a strong correlation between elements of a tricluster is insufficient to guarantee its statistical significance (deviation from expectations). Illustrating, small triclusters with good homogeneity are frequently observed, yet they generally occur by chance. Similarly, domain relevance (e.g., functional enrichment against knowledge bases [18]) is, alone, insufficient to this same end.

Henriques and Madeira [1] provide an entry point to recent research on tensor data analysis using triclustering, covering former studies

* Corresponding author at: Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal.
E-mail addresses: leonardoalexandre@tecnico.ulisboa.pt (L. Alexandre), rs.costa@fct.unl.pt (R.S. Costa), rmch@tecnico.ulisboa.pt (R. Henriques).

https://doi.org/10.1016/j.patcog.2023.110231
Received 28 June 2023; Received in revised form 14 October 2023; Accepted 26 December 2023
Available online 28 December 2023

Table 1
Well-established applications of triclustering in pattern recognition [19–21].

Domain	Data	Triclustering solutions
Biological	omic	Groups of genes in functional processes and pathways, co-regulated along time;
	pharmacological	Therapy-specific or dosing-discriminative regulatory responses;
	genomic	Conserved subsequences (alignments), binding site patterning in populations;
	networks	Modules of biological entities with interaction coherence along time;
	structural	Disease-specific glycoproteomic patterns; protein formation (residues-pos-time);
	others	Cross-species regulatory patterns; coherent pathology-specific phenotypes.
Medical	clinical records	Subsets of patients and symptoms with coherent health trajectories;
	multivariate signals	Coherent deviations from reference physiological patterning (risk groups).
	neuroimaging	Regions with properties of interest and characterization of functional connectivity;
Social	social media	Evolving communities of individuals with correlated activity and interaction;
	web	Users with coherent search and navigation behavior along time;
	e-commerce	Actionable local behaviors and preferences for recommendation systems;
	financial	Groups of stocks and investment decisions with similar indicators across time;
	geophysical	Spatially-dependent seismic activity; recurring meteorological phenomena.

that place statistical assessments [19,22–24]. Moise and Sander [22] placed a uniform null assumption to assess whether the volume of a given tricluster deviates from expectations under a Binomial test. Applicability is constrained to identically distributed variables. Sim et al. [23] proposed a correlation metric based on mutual information, termed CI, to identify triclusters with unexpectedly high CI (rarely co-occurring values), with the CI distribution being modeled under a gamma or Weibull distribution. Despite its relevance, statistical testing disregards temporal regularities and is computationally expensive, making it only viable for moderate-sized data. Mankad and Michailidis [24] explored tensor data resampling principles as a way of creating reference null data to statistically test triclusters. To determine the statistical significance of a candidate tricluster, its size is compared to the sizes of triclusters identified in randomized data (chance expectations). This approach disregards the distribution of values within a tricluster, being unable to model the likelihood of a specific pattern. Tchagang et al. [19] proposed statistical tests for order-preserving triclusters – patterns whose values follow specific ordering constraints within the selected observations –, assuming that time points are independent and variables are identically distributed according to a uniform distribution. A binomial distribution is applied to assess the statistical significance of an order-preserving slice. In this context, a statistically significant tricluster is one with statistically significant slices. Complementary statistical stances have been applied for different ends. Gutiérrez-Avilés and Rubio-Escudero [25,26] combined correlation coefficients with measures of cosine similarity among the slices of a tricluster (*observations* × *variables*, *observations* × *time points*, and *variables* × *time points*) and domain-specific functional annotations; while Biswal et al. [27] combined a classic homogeneity index (three-dimensional mean squared residue) with a statistical difference of a tricluster's values from the background (SDB) to ensure the presence of differential values against global regularities, and a functional enrichment score. Despite the relevance of these integrative statistics to promote homogeneity, reduce redundancy, and ensure domain relevance, they do not assess whether the found patterns deviate from expectations.

Triclustering approaches generally fail to assess or guarantee the statistical significance of the retrieved patterns (i.e. whether they do not occur by chance against a null data model) given the unique regularities of the available tensor data. At this light, we propose a solid statistical frame, termed TriSig, that expands well-established principles for assessing the statistical significance of biclustering solutions towards three-dimensional tensor data. In particular, TriSig is sensitive to: non-identically distributed variables, patterns with different homogeneity criteria, varying temporal regularities (e.g., contiguity and misalignments), and, consistently with Sim et al. [23], p-value corrections. Analogous to principles in information theory, the proposed statistical tests can be incorporated into an objective function to guide cluster formation [28,29]. In this context, this study presents the first comprehensive statistical frame to mitigate false positive discoveries from tensor data with varying properties.

The proposed methodology is assessed on four real-world case studies and synthetic tensor data. The application of triclustering algorithms with distinct search criteria behavior signaled vulnerabilities on their behavior and further showed how statistical significance can vary in accordance with the type and size of patterns, background data regularities, dependency assumptions, and selected corrections. These later knowledge can be used to represent surfaces to discriminate spurious discoveries from statistically significant ones.

The manuscript is organized as follows: the remainder of this section provides the essential background. Section 2 introduces the proposed methodology, while Section 3 assesses the methodology against alternative statistical stances using real-world and synthetic data. Concluding remarks and implications are finally drawn.

1.1. Background

Let a two-dimensional dataset (matrix data), **A**, be defined by n observations ($X = \{x_1, \dots, x_n\}$) and m variables ($Y = \{y_1, \dots, y_m\}$), with $n \times m$ elements a_{ij}. *Clustering* and *biclustering* tasks aim to extract *clusters* $I_i \subseteq X$ or *biclusters* $B_k = (I_k, J_k)$, where a given subset of observations, $I_k \subseteq X$, is correlated on a subset of variables, $J_k \subseteq Y$. **Homogeneity** criteria, commonly guaranteed through the use of a merit function, guides the formation of (bi)clusters [30]. Given a bicluster, its values $a_{ij} = c_j + \gamma_i + \eta_{ij}$ can be described by value expectations c_j, adjustments γ_i, and noise η_{ij}, with a **bicluster pattern** φ_B being the ordered set of expected values in the absence of adjustments and noise, $\varphi_B = \{c_j | y_j \in J\}$. In addition to homogeneity criteria, **statistical significance** criteria [31] guarantee that the probability of a bicluster's occurrence (against a null model) deviates from expectations. Furthermore, **dissimilarity** criteria [32] can be placed to further guarantee the absence of redundant biclusters.

A three-dimensional dataset (*three-way tensor data*), **A**, is defined by n observations $X = \{x_1, \dots, x_n\}$, m variables $Y = \{y_1, \dots, y_m\}$, and p contexts $Z = \{z_1, \dots, z_p\}$. Elements a_{ijk} relate observation x_i, attribute y_j, and context z_k. Identical to two-dimensional data, three-way tensor data can be *real-valued* ($a_{ijk} \in \mathbb{R}$), symbolic ($a_{ijk} \in \Sigma$, where Σ is a set of nominal or ordinal symbols), integer ($a_{ijk} \in \mathbb{Z}$), or *non-identically distributed* ($a_{ijk} \in \mathcal{A}_j$, where $\mathcal{A}_j$ is the domain of y_j's variable), and further contain missing elements, $a_{ijk} \in \mathcal{A}_j \cup \emptyset$. When the context dimension corresponds to a set of time points ($Z = \{t_1, t_2, \dots, t_p\}$), we are in the presence of temporal three-dimensional dataset, also referred as m-order *multivariate time series data*.

Given a three-way tensor dataset, **A**, with n observations, m variables, and p contexts/time points, a **tricluster** $T = (I, J, K)$ is a subspace of the original space, where $I \subseteq X$, $J \subseteq Y$, and $K \subseteq Z$ are subsets of observations, variables, and contexts/time points, respectively [1]. The **triclustering task** aims to find a set of triclusters $\mathcal{T} =$

$\{T_1, \ldots, T_l\}$ such that each tricluster T_i satisfies specific criteria of homogeneity, dissimilarity, and statistical significance. Let each element of a tricluster, a_{ijk}, be described by a base value c_{jk}, observation-specific adjustments γ_i and noise η_{ijk}. The tricluster **pattern**, φ_T, is an ordered set of value expectations along the subset of variables and context dimensions in the absence of adjustments and noise: $\varphi_T = \{c_{jk} | y_j \in J, z_k \in K, c_{jk} \in Y_k\}$.

Homogeneity criteria determine the structure, coherence, and quality of a triclustering solution [1], where: (1) the *structure* is described by the number, size, shape, and position of triclusters, (2) the *coherence* of a tricluster is defined by the observed correlation of values (coherence assumption) and the allowed deviation from expectations (coherence strength), and (3) the *quality* of a tricluster is defined by the type and amount of tolerated noise. A tricluster has **constant coherence** when the subspace exhibits constant (for symbolic data) or approximately constant (real-valued data) values. **Dissimilarity** criteria guarantee that any tricluster similar to another tricluster with higher priority is removed from $\mathcal{T}$ and (possibly) used to refine similar triclusters in $\mathcal{T}$. In this context, a tricluster $T = (I, J, K)$ is *maximal* if and only if there is no other tricluster $(\mathbf{I}', \mathbf{J}', \mathbf{K}')$ such that $I \subseteq I' \wedge J \subseteq J' \wedge K \subseteq K'$ satisfying the given criteria. Finally, and foremost, **statistical significance criteria** should be placed to guarantee that the probability of observing a given tricluster against a null data model is unexpectedly low.

2. Methodology

The proposed methodology offers a statistical frame to assess the likelihood of observing a given tricluster against sound null data models, accommodating the necessary principles to model local regularities observed in tensors and non-identically distributed multivariate time series data. Our methodology expands upon two robust statistical stances designed for biclustering solutions: (1) the BSig tests proposed by Henriques and Madeira [31] where binomial testing is flexibly applied considering the underlying pattern types, data regularities (e.g., empirical and theoretical stances for both real-valued and categorical data), and multi-hypothesis corrections; and (2) CCC-Biclustering algorithm, proposed by Madeira et al. [33] to handle time series data, including Markovian assumptions and applicable corrections for temporal misalignments [34].

Given a tensor $\mathbf{A}$, to estimate the probability of a tricluster's pattern (slice), p_{φ_T}, to occur against null expectations, we must make assumptions regarding local conditional dependencies among variables and contexts – mutually dependent (MD), mutually independent (MI), temporal contiguity (TC) – and variable distributions – identical versus non-identical. To this end, it is important to understand the type of data at hands. Table A.6 provides illustrative data domains, providing a view on which of the aforementioned assumptions can be placed to adequately create a null data model. For example, local dependencies between co-expressed genes in omic data domains are commonly placed due to their co-regulation in various pathways, affecting the probability calculus [35]. Similarly, classic variable independence assumptions in null models can be replaced with local dependencies in the context of weather pattern prediction to better reflect the strength of associations between thermodynamic and environmental variables [36]. This type of analysis is crucial to correctly model null expectations as they determine the final significance estimates.

2.1. Tricluster pattern's probability

Given a tricluster T with pattern φ_T, if we assume that variables are mutually independent and that contexts are also mutually independent, then the probability of the tricluster's pattern occurring is given by

$$p_{\varphi_T} = \prod_{z_k \in K} p_{\varphi_{B_k}} = \prod_{z_k \in K} \prod_{y_j \in J} P(c_{jk}), \tag{1}$$

where K comprises the contexts in the tricluster, J is the set of variables in the tricluster, $p_{\varphi_{B_k}}$ is the probability of a bicluster's pattern φ_{B_k} in context z_k to occur against null expectations, $P(c_{jk})$ is an estimate of the probability of value c_{jk} occurrence, in the absence of adjustments and noise. As both variables and contexts are mutually independent, the probability of the value c_{jk} is approximated by the available n observations measured by variable y_j in context z_k.

If we consider variables to be mutually dependent and contexts to be mutually independent, then a tricluster pattern's probability is given by

$$p_{\varphi_T} = \prod_{z_k \in K} p_{\varphi_{B_k}} = \prod_{z_k \in K} P(\bigwedge_{y_j \in J} c_{jk}). \tag{2}$$

When contexts are mutually dependent, the probability of value c_{jk} is approximated by the set of measurements from y_j across all contexts, Z. Additionally, as we need to account for the pattern to occur in any context of Z, then

$$p_{\varphi_T} = \binom{|Z|}{|K|} \prod_{z_k \in K} p_{\varphi_{B_k}} = \binom{|Z|}{|K|} \prod_{z_k \in K} \prod_{y_j \in J} P(c_{jk}), \tag{3}$$

when variables are mutually independent, and

$$p_{\varphi_T} = \binom{|Z|}{|K|} \prod_{z_k \in K} p_{\varphi_{B_k}} = \binom{|Z|}{|K|} \prod_{z_k \in K} P(\bigwedge_{y_j \in J} c_{jk}), \tag{4}$$

when variables are mutually dependent. To further clarify the calculus, an example of the above formulas, using a dummy data example, is given in Fig. 1.

In the presence of temporal data, where Z represents the set of time points and K is a contiguous subset of Z, then the null probability of a tricluster's pattern, p_{φ_T}, can be modeled by a first-order Markov chain. Instead of accounting for the possibility of the pattern occurring in any context of Z, we need to accommodate the possibility for temporal misalignments [34]. To this end, p_{φ_T} is given by

$$p_{\varphi_T} = (|Z| - |K| + 1) \times p_{\varphi_{B_{k_1}}} \times \prod_{z_k \in K \setminus z_{k_1}} p_{\varphi_{B_k} | \varphi_{B_{k-1}}} \tag{5}$$

where $p_{\varphi_{B_{k_1}}}$ is given by $\prod_{y_j \in J} P(c_{jk_1})$ if variables are assumed to be mutually independent or $P(\bigwedge_{y_j \in J} c_{jk_1})$ if we assume mutual dependence, and $p_{\varphi_{B_k} | \varphi_{B_{k-1}}}$ represents the probability of the pattern transitioning from context z_{k-1} to context z_k. The probability of transitioning depends on whether we assume variables to be mutually dependent or independent. Considering the Bayes rule, when assuming variables to be mutually independent, then

$$p_{\varphi_{B_k} | \varphi_{B_{k-1}}} = \prod_{y_j \in J} \frac{P(c_{jk}, c_{j(k-1)})}{P(c_{j(k-1)})}, \tag{6}$$

and, when assuming mutually dependent variables, then

$$p_{\varphi_{B_k} | \varphi_{B_{k-1}}} = \frac{P(\bigwedge_{y_j \in J} c_{jk}, c_{j(k-1)})}{P(\bigwedge_{y_j \in J} c_{j(k-1)})}, \tag{7}$$

Fig. 1(d) presents an example of the aforesaid calculus when in the presence of mutually independent variables.

The introduced probabilistic stances can be considered for categorical and real-valued variables, and either rely on empirical or theoretical distributions. The adequacy of the approximated distributions can be assessed using χ^2 goodness-of-fit test for categorical data and Kolmogorov–Smirnov goodness-of-fit for real-valued data. For details on how to circumvent point estimate problems in real-valued data, we redirect the reader to [31].

2.2. Tricluster's statistical significance

Given p_{φ_T}, binomial tails can be used to robustly compute the probability of a tricluster $T = (I, J, K)$ with pattern φ_T to occur for

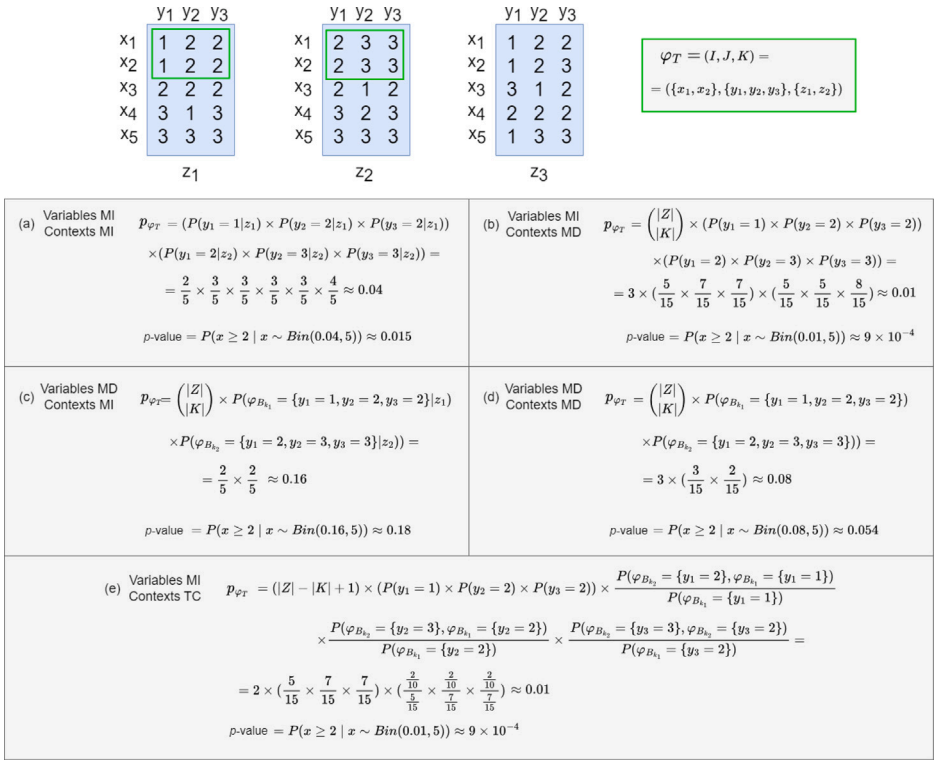


Fig. 1. Illustrative example of statistical significance calculus under different data assumptions.

$|I|$ or more observations [31]. Accordingly,

$$p\text{-value} = P(x \geq |I| \mid x \sim Bin(p_{\varphi_T}, |X|)) = \sum_{x=|I|}^{|X|} \binom{|X|}{x} p_{\varphi_T}^{x} (1 - p_{\varphi_T})^{|X|-x}, \tag{8}$$

where $Bin(p_{\varphi_T}, |X|)$ is the null binomial distribution, p_{φ_T} the null pattern probability, $|X|$ the total number of observations, and $|I|$ the number of successes.

When variables are identically distributed, the p-value can be further corrected when considering that the observed pattern can equally span alternative variables [31], i.e.,

$$p\text{-value} = \binom{|Y|}{|J|} \times P(x \geq |I| \mid x \sim Bin(p_{\varphi_T}, |X|)). \tag{9}$$

Irrespectively of the above relaxation, further corrections should be placed for multiple hypotheses (patterns). To this end, the Benjamini–Hochberg procedure [37] is suggested to decrease the false discovery rate. This correction tackles the problems faced by more conservative corrections, such as Bonferroni, in the presence of a high number of hypotheses.

Instantiating the introduced methodology, Table 2 shows the minimum number of observations required for a tricluster's pattern φ_T to be statistically significant ($\alpha < 0.01$). In particular, given a three-way data with uniformly distributed variables, this analysis shows how the minimum number of observations varies for: (i) patterns with varying number of variables $|J|$ and contexts $|K|$, (ii) categorical variables with varying cardinality ($|L|$), (iii) data with varying dependency assumptions, and (iv) the presence of multi-hypotheses corrections. We confirm that "Variables MI, Contexts MD" assumption is stricter than "Variables MI, Contexts TC", generally requiring fewer tricluster's observations to guarantee statistical significance. The pattern size ($|J|$ and $|K|$) heavily influence statistical significance, with $|K|$ being particularly relevant under the latter assumption.

3. Computational and experimental setup

Triclustering algorithms. Triclustering algorithms can be categorized according to specific characteristics, including the underlying search procedure, targeted homogeneity, allowed data types (e.g., real-valued, ordinal, binary tensor data), amongst others [1]. To fully explore the proposed solution, we selected three triclustering algorithms representative of three major search criteria – *heuristic*-based (greedy) searches, *exhaustive* searches based on formal concept analysis, and *quasi-exhaustive* searches grounded on slice-based consensus. Accordingly, the selected algorithms are (1) δ-Trimax [38], a greedy iterative search algorithm that handles real-valued tensor; (2) TRIAS algorithm [39], an exhaustive algorithm that only handles binary data; and (3) modified version of Zaki *TriCluster* algorithm [40], a consensus-based algorithm with a modification that restricts the patterns to be contiguous along the context axis.

Datasets. To validate the proposed methodology, three publically available **case studies** are selected: *glycine* data [41], *gene expression* profile [42], and *penicillin batch fermentation* [43,44]. Table 3 provides a brief description of each dataset and the applied preprocessing.

To complement the analysis, we generated **synthetic data** using G-Tric, a synthetic data generator proposed by Lobo et al. [45], able to generate tensors with planted triclusters. The authors provide a set of reference properties based on the type of targeted data such as gene expression, financial, fMRI, social, and georeferenced time-series. With this in mind, the generated datasets contain 1000 observations, 50 variables, 50 contexts, and two types of background distribution – uniform and Gaussian ($\mu = 0.0, = 30$) –, and 5 triclusters planted per dataset. The artificial variables are ordinal with cardinality $|L| \in \{3, 5, 7\}$. The planted triclusters had the following additional properties: (1) the number of observations between 50 and 500, (2) the number of variables between 2 and 4, (3) the number of contexts between 2 and 4, and (4) patterns could be contiguous or non-contiguous across contexts.

Table 2
Given uniformly distributed three-way data ($|X|$=1000, $|Y|$=50, $|Z|$=50), the minimum number of observations of a tricluster, n_{min}, required to be statistically significant (p-value < 0.01) is assessed. In particular, we measure the impact of pattern size (number of variables $|J|$ and contexts, $|K|$), cardinality ($|L|$), and corrections. Two major local assumptions are considered to this end, "Variables MI, Contexts MD" where $P(\varphi_T) = \frac{1}{|L|}^{|J|\times|K|} \times \binom{|Z|}{|K|}$ and "Variables MI, Contexts TC" where $P(\varphi_T) = \frac{1}{|L|}^{|J|\times|K|} \times (|Z| - |K| + 1)$.

$\|X\| = 1000, \|Y\| = 50, \|Z\| = 50$ (variables with uniform distribution)									
		Variables MI, Contexts MD				Variables MI, Contexts TC			
	$\|J\|$	2	2	2	2	2	2	2	2
	$\|K\|$	2	3	4	5	2	3	4	5
$\|L\|$=3	n_{min}	–	–	–	–	641	85	14	3
	n_{min} (corrected)	–	–	–	–	671	102	21	7
	$P(\varphi_T)$	15.12	26.88	35.10	35.88	0.60	0.07	7.16×10^{-3}	7.79×10^{-4}
$\|L\|$=5	n_{min}	–	–	626	248	99	8	1	1
	n_{min} (corrected)	–	–	656	275	117	13	3	2
	$P(\varphi_T)$	1.96	1.25	0.59	0.22	0.08	3.07×10^{-3}	1.20×10^{-4}	4.71×10^{-6}
	$\|J\|$	3	3	3	3	3	3	3	3
	$\|K\|$	2	3	4	5	2	3	4	5
$\|L\|$=3	n_{min}	–	–	470	174	86	7	1	1
	n_{min} (corrected)	–	–	510	205	109	13	4	2
	$P(\varphi_T)$	1.68	0.99	0.43	0.14	0.07	2.43×10^{-3}	8.84×10^{-5}	3.20×10^{-6}
$\|L\|$=5	n_{min}	99	18	4	1	8	1	1	1
	n_{min} (corrected)	123	29	9	4	15	3	1	1
	$P(\varphi_T)$	0.08	0.01	9.43×10^{-4}	6.94×10^{-5}	3.13×10^{-3}	2.45×10^{-5}	1.92×10^{-7}	1.51×10^{-9}
	$\|J\|$	4	4	4	4	4	4	4	4
	$\|K\|$	2	3	4	5	2	3	4	5
$\|L\|$=3	n_{min}	216	51	11	3	14	1	1	1
	n_{min} (corrected)	255	73	22	8	26	5	2	1
	$P(\varphi_T)$	0.18	0.04	5.35×10^{-3}	6.07×10^{-4}	7.46×10^{-3}	9.03×10^{-5}	1.09×10^{-6}	1.31×10^{-8}
$\|L\|$=5	n_{min}	8	1	1	1	1	1	1	1
	n_{min} (corrected)	16	4	1	1	5	1	1	1
	$P(\varphi_T)$	3.14×10^{-3}	8.03×10^{-5}	1.51×10^{-6}	2.22×10^{-6}	1.25×10^{-4}	1.97×10^{-7}	3.08×10^{-10}	4.82×10^{-13}
	$\|J\|$	5	5	5	5	5	5	5	5
	$\|K\|$	2	3	4	5	2	3	4	5
$\|L\|$=3	n_{min}	32	5	1	1	4	1	1	1
	n_{min} (corrected)	51	12	5	2	10	3	1	1
	$P(\varphi_T)$	0.20	1.36×10^{-3}	6.60×10^{-5}	2.50×10^{-6}	8.29×10^{-4}	3.34×10^{-6}	1.34×10^{-8}	5.42×10^{-11}
$\|L\|$=5	n_{min}	1	1	1	1	1	1	1	1
	n_{min} (corrected)	6	2	1	1	3	1	1	1
	$P(\varphi_T)$	1.25×10^{-4}	6.42×10^{-7}	2.41×10^{-9}	7.11×10^{-12}	5.02×10^{-6}	1.57×10^{-9}	4.92×10^{-13}	1.54×10^{-16}

Table 3
Summary of the selected case studies, including the size and domain of each dimension (e.g., batches, units, genes), and the applied preprocessing.

Datasets	Description	Preprocessing
glycine (subjects x units x time) (23 x 467 x 6)	A metabolomics dataset from urinary samples analyzed by nuclear magnetic resonance spectroscopy giving a total of 467 integrated units per NMR spectrum for each observation. In this study, they used commercial benzoic acid containing flavored water as a vehicle for designed interventions.	– $\|L\| = 5$ (ordinal categories)
gene expression (samples x genes x time) (15 x 500 x 4)	Liver tissue transcriptomics (45101 genes) from male mice in response to the application of tetrachlorodibenzodioxin (TCDD) administered orally at doses of 0, 1, 3, 10 and 30 μg/kg. The liver was sampled 2, 4, 8 and 24 h after administration.	– selection of top 500 genes with highest variability on $X \times Z$; – $\|L\| = 3$
penicillin batch fermentation (batches x variables x time) (100 x 16 x 25)	An advanced mathematical model of a 100,000 litre penicillin fermentation system (bioprocess data). It contains 100 batches, where batches 1–30 are controlled by recipe-driven approach, 31–60 are controlled by operators, 61–90 are controlled by an Advanced Process Control (APC) solution using the Raman spectroscopy, and 91–100 contain faults resulting in process deviations. Each batch is a multivariate time series where all available processes, model inputs, and Raman spectroscopy measurements are available. There is not a fixed number of time points per batch.	– removed Raman spectroscopy; – removed uninformative variables; – Piecewise Aggregate Approximation (PAA); (25 time points) – $\|L\| = 7$

Pattern evaluation. In synthetic data, there is no prior knowledge on variable dependencies and all variables are sampled from either a Gaussian or uniform distribution. To this end, for the purpose of testing the proposed statistical frame, we assume variables to be identically distributed and two distinct null assumptions: (1) mutually independent variables, and (2) mutually dependent variables. Regarding context dependency, we assume the following cases: (1) contexts to be mutually independent, (2) contexts to be mutually dependent, and (3) contexts to represent a set of time points.

The selected real-world case studies are three-way time series data. This means that patterns that exhibit contiguity across this dimension can be evaluated under a Markovian assumption, such as a first-order Markov chain. Out of the three algorithms, only the modified Zaki algorithm [40] extracts patterns contiguous across time. As such, the patterns extracted using the modified Zaki algorithm will be analyzed assuming a first-order Markov chain, whilst the patterns extracted using TRIAS and δ-Trimax will be analyzed assuming contexts to be mutually dependent. The variables in the *penicillin batch fermentation* and *glycine*

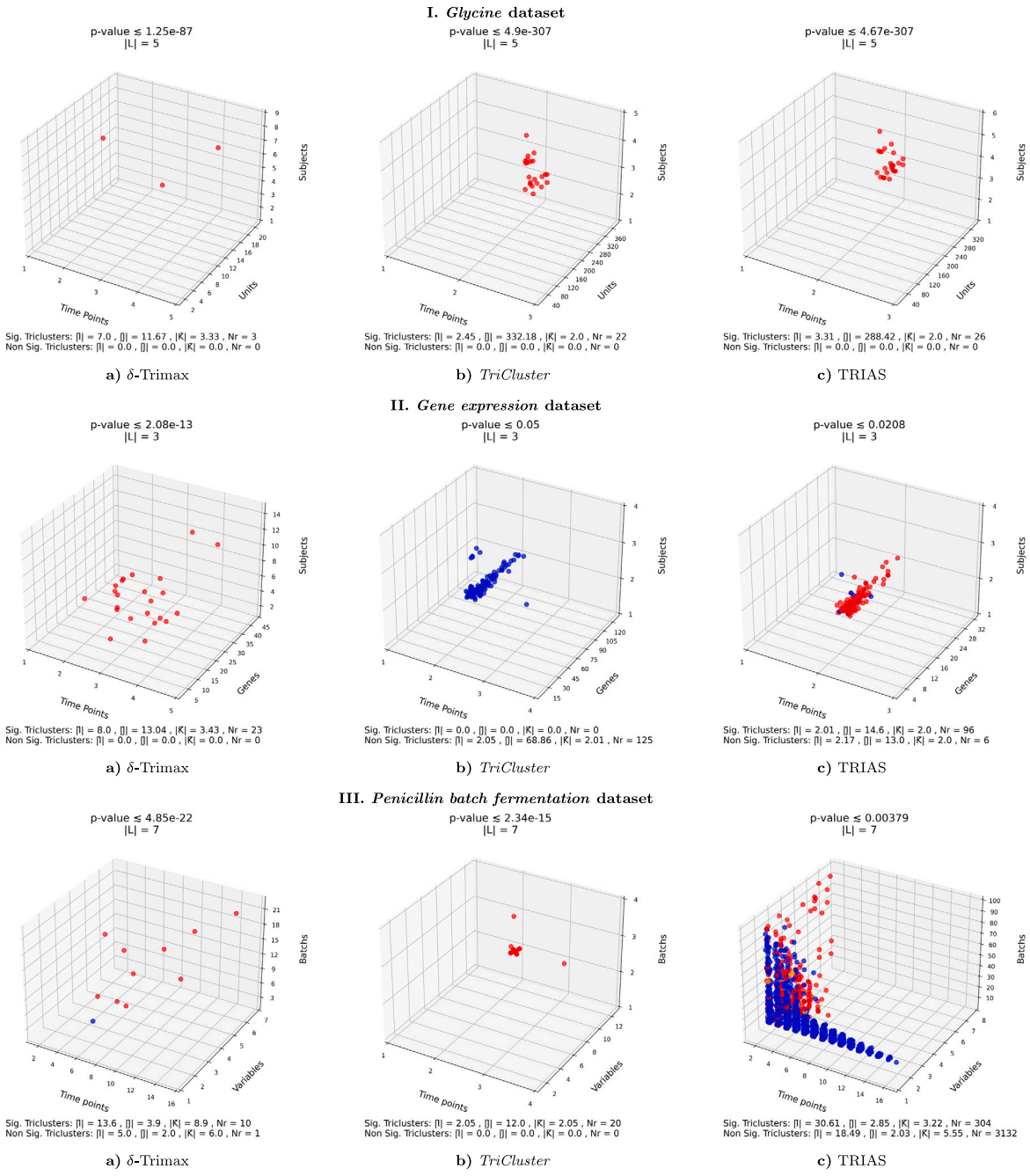


Fig. 2. Solutions $\bar{I}$ returned from triclustering algorithms (δ-Trimax, *TriCluster* and TRIAS) applied to three real-world datasets. Each dot on the scatter plot represents a tricluster where its position is defined by the number of observations ($|I|$), variables ($|J|$), and time points ($|K|$). The dots are colored red if they are below a statistical significance threshold defined by the Benjamini–Hochberg procedure, colored orange if they are above the threshold but below 0.05, and colored blue otherwise. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cases are assumed to be independent in the null data model, while locally dependent for the *gene expression* case.

Baselines. We further conduct a comparative analysis against two distinct methodologies described in [22] and [23], quantifying and discussing the differences amongst the observed statistical measurements.

4. Results and discussion

The proposed TriSig methodology is first applied in real-world case studies and synthetic data to assess its validity and implications. We further compare TriSig against alternative statistical methodologies to stress the relevance of placing proper null model assumptions.

Real-world case studies

Consider the *glycine* dataset, with results presented in Fig. 2-I. All triclustering algorithms extracted a compact set of patterns. The δ-Trimax extracted a total of 3 patterns, with an average number of 3 subjects (13%), 11 units (2%), and 3 time points (50%). *TriCluster* and TRIAS extracted slightly more patterns, 22 and 26 accordingly, but with a lower average of subjects, between 2 and 5 subjects (5%–20%). Despite the different placed assumptions, the patterns extracted by all algorithms generally yield statistical significance (low p-values). The Benjamini–Hochberg procedure sets the statistical significance thresholds between 4.9×10^{-307} and 1.25×10^{-87}. This can be explained by the

Table 4
Statistics of the triclusters presented in Fig. 3 when variables either follow a uniform or Gaussian distribution. Column "$(|I|, |J|, |K|)$" presents details pertaining to the volume of the triclusters. Column "$p_{\varphi_{B_k}}$" presents intermediate statistics necessary to calculate p_{φ_T}. Column "p-value" and "p-value (with correction)" contain the p-values obtained using the binomial test of significance, with and without the proposed correction.

Triclusters $(\|I\|, \|J\|, \|K\|)$	Synthetic data ($\|X\| = 1000$, $\|Y\| = 50$, $\|Z\| = 50$) – Uniform distribution		
	$p_{\varphi_T} = \binom{Z}{k} \prod_{k \in K} p_{\varphi_{B_k}}$	p-value	p-value (with correction)
(55, 3, 2)	$p_{\varphi_{B_{k_1}}} \approx 5.2 \times 10^{-3}$	2.9×10^{-4}	1.0
(484, 2, 2)	$p_{\varphi_{B_{k_1}}} \approx 0.04$	1.0	1.0
(249, 3, 2)	$p_{\varphi_{B_{k_1}}} \approx 0.01$	8.9×10^{-6}	0.17
(327, 3, 2)	$p_{\varphi_{B_{k_1}}} \approx 0.02$	0.13	1.0
(85, 2, 3)	$p_{\varphi_{B_{k_1}}} \approx 0.03$	1.0	1.0

Triclusters $(\|I\|, \|J\|, \|K\|)$	Synthetic data ($\|X\| = 1000$, $\|Y\| = 50$, $\|Z\| = 50$) – Gaussian distribution		
	$p_{\varphi_T} = (\|Z\| - \|K\| + 1) \times p_{\varphi_{B_{k_1}}} \times \prod_{k \in K \setminus k_1} p_{\varphi_{B_k} \mid \varphi_{B_{k-1}}}$	p-value	p-value (with correction)
(368, 3, 3)	$p_{\varphi_{B_{k_1}}} = 0.11$, $p_{\varphi_{B_{k_2}} \mid \varphi_{B_{k_1}}} = 0.22$, $p_{\varphi_{B_{k_3}} \mid \varphi_{B_{k_2}}} = 0.22$	2.48×10^{-14}	4.87×10^{-10}
(498, 2, 3)	$p_{\varphi_{B_{k_1}}} = 0.24$, $p_{\varphi_{B_{k_2}} \mid \varphi_{B_{k_1}}} = 0.29$, $p_{\varphi_{B_{k_3}} \mid \varphi_{B_{k_2}}} = 0.29$	1.0	1.0
(334, 3, 3)	$p_{\varphi_{B_{k_1}}} = 0.12$, $p_{\varphi_{B_{k_2}} \mid \varphi_{B_{k_1}}} = 0.22$, $p_{\varphi_{B_{k_3}} \mid \varphi_{B_{k_2}}} = 0.22$	0.003	1.0
(275, 3, 3)	$p_{\varphi_{B_{k_1}}} = 0.11$, $p_{\varphi_{B_{k_2}} \mid \varphi_{B_{k_1}}} = 0.20$, $p_{\varphi_{B_{k_3}} \mid \varphi_{B_{k_2}}} = 0.20$	7.7×10^{-7}	0.02
(229, 3, 2)	$p_{\varphi_{B_{k_1}}} = 0.12$, $p_{\varphi_{B_{k_2}} \mid \varphi_{B_{k_1}}} = 0.17$	1.0	1.0

imposed dependency assumption along the units' dimension, associated with low co-occurrence probabilities.

Essential results for the transcriptional drug response data are presented in Fig. 2-II. In this case, δ-Trimax extracted a total of 23 patterns, while *TriCluster* and TRIAS extracted a total of 125 and 102 patterns respectively. Patterns extracted by δ-Trimax have an average number of 8 subjects (67%), 13 genes (3%), and 3 time points (75%). *TriCluster* and TRIAS have a lower number of subjects (16%), 68 genes (13%) in *TriCluster* and 14 genes (3%) in TRIAS, and 2 time points (50%). Contrary to the glycine case study, due to the more conservative null model assumption, all patterns extracted by *TriCluster* have a p-value above the imposed significance threshold. Similarly, some patterns extracted by TRIAS also fail to meet the imposed significance threshold.

Results for the industrial batches' study are presented in Fig. 2-III. Both δ-Trimax and the *TriCluster* extracted a small number of patterns (11 and 20), whilst TRIAS retrieved over 3400 patterns. For the *TriCluster* algorithm, all patterns yield statistical significance on this case study. The patterns extracted by *TriCluster* occur on average in 2 batches (2%), 12 variables (75%), and 2 time points (8%), meaning that despite their low frequency (low number of batches), the retrieved patterns have enough variables and time points to be marked as statistically significant. Considering δ-Trimax results, the number of variables play an important role to determine whether patterns are significant (3 or more variables) or non-significant. Patterns with statistical significance contain on average 14 batches (13%), 4 variables (25%) and 9 time points (36%). Contrary to previous case studies, TRIAS extracted 9% of patterns with statistical significance and 3132 spurious patterns (91%). As depicted in Fig. 2-III, most false discoveries show a lower number of variables (between 2 and 3) against statistically significant patterns. In this case, we can observe that the number of time points does not strongly condition the pattern's statistical significance given the presence of spurious patterns with up to 16 time points.

Synthetic data

What makes a tricluster statistically significant? To answer this research question, synthetic data with planted triclustering solutions are used to statistically assess triclusters with varying properties.

Consider the results for the synthetic data presented in Fig. 3. Assuming variables and contexts to be mutually dependent, we selected two cases: (1) variables following a uniform distribution and planted triclusters with non-contiguous contexts, and (2) variables following a Gaussian distribution and triclusters with contiguous contexts (time points). The pattern characteristics, intermediate statistics, and p-values are presented in Table 4. By inspecting the aforesaid figure and table, we observe that the randomly planted triclusters with statistical significance are not only conditioned by the number of observations, variables, and contexts they occur, but as well by the probability of the tricluster's pattern, φ_T. In this case, with mutually dependent contexts and no context contiguity, the joint probability of all slice patterns, φ_B, contained in φ_T is balanced by $\binom{Z}{k}$. If the enclosed patterns φ_B are not sufficiently improbable, the associated triclusters are generally spurious. We observe this in triclusters ($|I| = 55, |J| = 3, |K| = 2$) and ($|I| = 85, |J| = 2, |K| = 3$) since the pattern is not rare enough to be considered statistically significant. Considering the Gaussian setting, where contexts are mutually dependent and triclusters contiguous across contexts, the joint probability between $\varphi_{B_{k_1}}$ and φ_T transitions, and a more relaxed statistic ($|Z| - |K| + 1$), allow more probable patterns to have statistical significance. Finally, the changes observed under the p-value correction suggest the importance of this procedure to effectively mitigate possible false positives.

Comparative analysis of methodologies

Comparative results produced by our methodology (TriSig) against the statistical methodologies proposed by Sim et al. [23] (SimSig) and Moise and Sander [22] (MoiseSig) are provided in Table 5 and Fig. 4.

Considering the patterns extracted by δ-Trimax algorithm, both TriSig and MoiseSig produced similar assessments with regards to the number of statistically significant patterns, diverging only for the *gene expression* dataset. In particular, patterns assessed by MoiseSig presents a lower p-value variability ([Q1,Q3]). In contrast, SimSig diverges in their assessments, classifying most patterns as spurious discoveries. A similar appraisal can be acquired when considering the patterns extracted by the *TriCluster* algorithm. Both TriSig and MoiseSig classify all patterns as statistically significant (or non-significant for the case of *gene expression* dataset), whilst SimSig classifies most patterns as false discoveries. Considering the patterns extracted by TRIAS algorithm, MoiseSig and TriSig produced similar assessments for both *glycine* and *gene expression* datasets. Similarly to previous algorithms, MoiseSig shows lower p-value variability and SimSig produces discordant assessments. In the context of the top 300 patterns produced for the *penicillin batch fermentation* dataset, all three approaches show diverging assessments. MoiseSig classifies every pattern as statistically

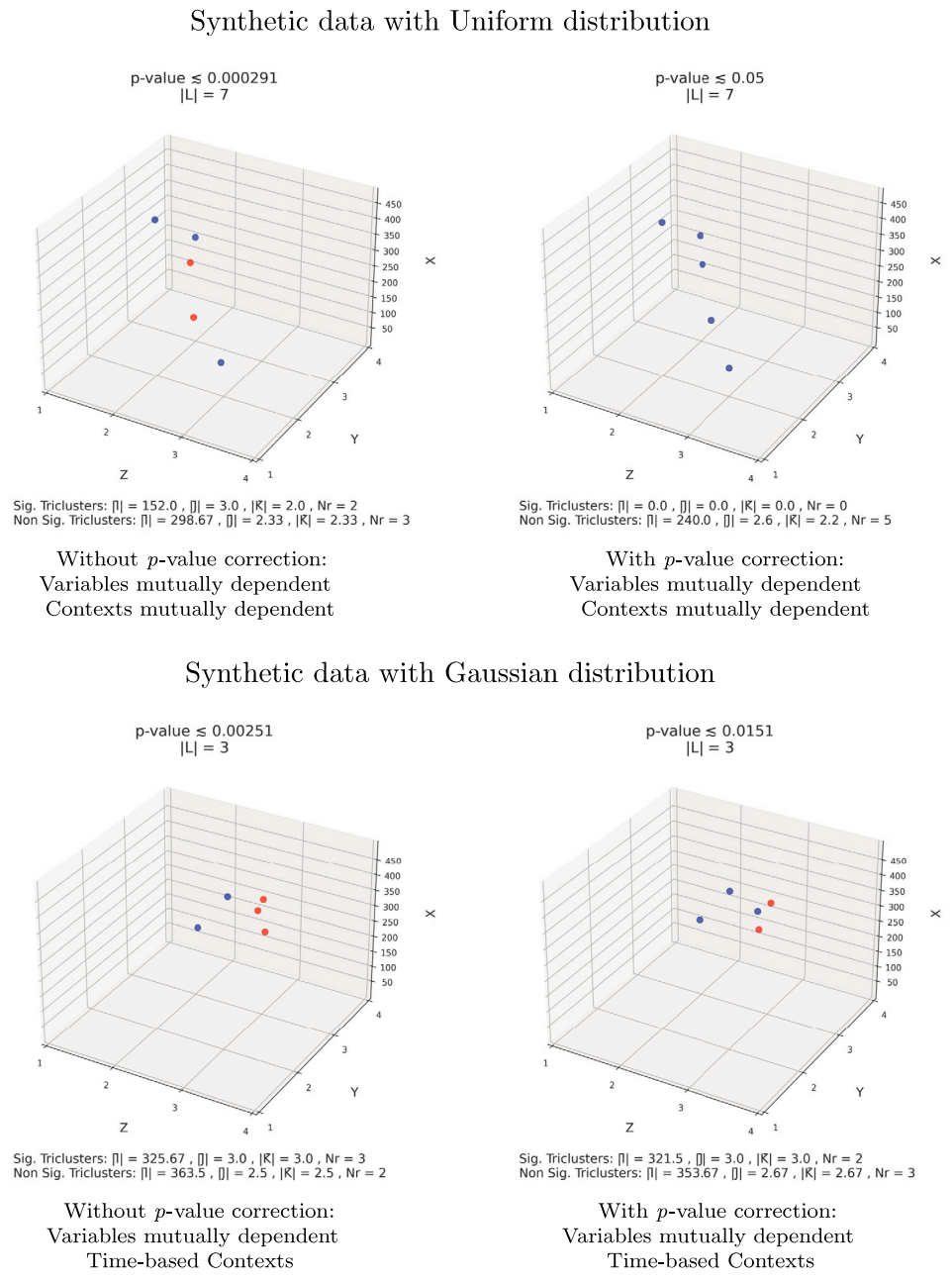


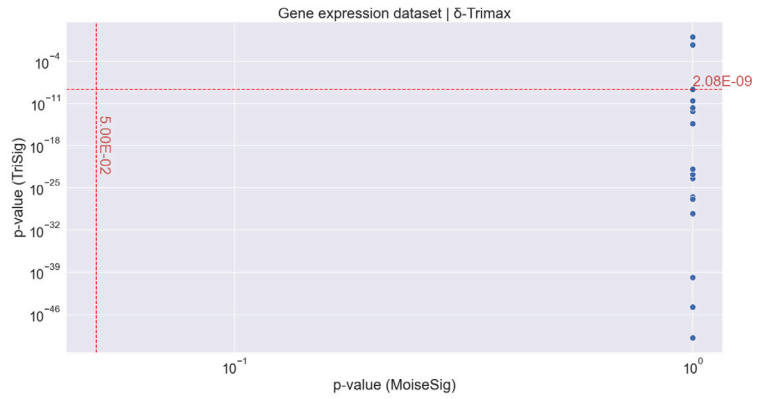
Fig. 3. Results for the planted triclusters in synthetic data when variables either follow an Uniform or Gaussian distribution, considering varying dependency and time contiguity assumptions. Each dot on the scatter plot represents a tricluster. The position of the tricluster is defined by its number of observations ($|I|$), variables ($|J|$), and contexts ($|K|$). The dots are colored red if they are below the statistical significance threshold defined by the Benjamini–Hochberg procedure, colored orange if they are above the threshold but are below 0.05, and colored blue otherwise. In the first column of scatter plots p-value correction is not applied, in the second column p-value correction is applied. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

significant (Q1 = 2.23×10^{-308}, Q3 = 1.81×10^{-91}), while SimSig classifies approximately half of the patterns as significant with a narrower distribution of p-values (Q1 = 2.25×10^{-9}, Q3 = 4.59×10^{-2}). Finally, TriSig (Q1 = 1, Q3 = 1) classifies approximately a fifth of the patterns as statistically significant.

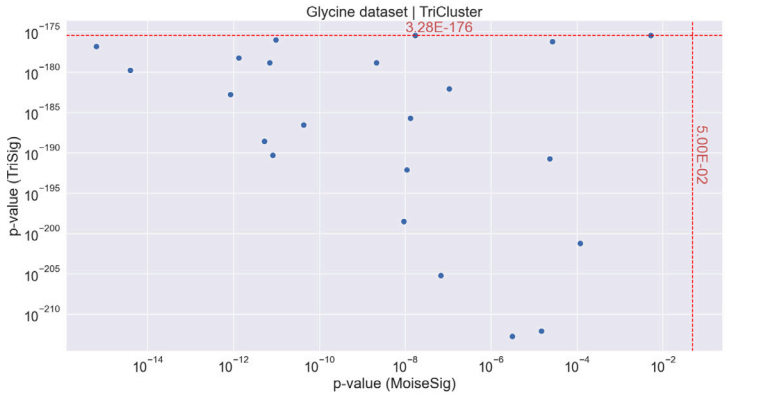
To complement Table 5, Fig. 4 compares the statistical significance of the retrieved triclusters from a dataset produced from different statistical methodologies – TriSig against MoiseSig or SimSig – by plotting the triclusters' p-values in relation to the reference significance thresholds. Fig. 4a presents a case where MoiseSig and TriSig diverge in their statistical significance assessment. We observe that MoiseSig considers all retrieved patterns as false discoveries, whilst TriSig considered only a fraction of them (patterns above the dotted red line). Fig. 4b presents a case where MoiseSig and TriSig consider most patterns to be statistically significant, although heightened differences in p-values are observed, driven by distinct null assumptions. Finally, Fig. 4c presents results on penicillin batch fermentation where TriSig and SimSig consider only a fraction of the patterns to be statistically significant with small-to-moderate agreement (where concordance is observed on the lower left and upper right quadrants).

5. Conclusion

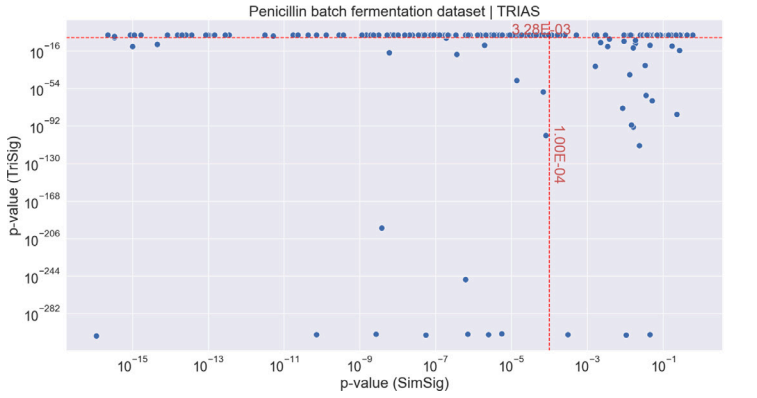
This work proposes a novel methodology to rigorously assess the statistical significance of patterns extracted from tensor data using triclustering. The methodology provides statistical principles to model



a) patterns extracted by δ-Trimax over the *gene expression* dataset, and when comparing TriSig against MoiseSig



b) patterns extracted by TriCluster over the *glycine* dataset, and when comparing TriSig against MoiseSig



c) patterns extracted by TRIAS over the *penicillin batch fermentation* dataset, and when comparing TriSig against SimSig

Fig. 4. Results produced from comparing the statistical significance of retrieved triclusters using different algorithms and datasets. In each scatter plot, a dot represents a tricluster located by the *p*-values obtained using TriSig and either MoiseSig or SimSig.

local regularities of tensor data with varying variable domains, dependencies and temporal dynamics, as well as subsequent statistical testing against null models. To assess its validity we: (1) conducted a theoretical analysis to identify the minimum number of observations necessary to guarantee the statistical significance of a tricluster as a function of the pattern length, number of time points, and the above-stated regularities; (2) applied our methodology over distinct real-world case studies using distinct triclustering algorithms (heuristic, exhaustive, quasi-exhaustive approaches), which revealed unique behaviors and vulnerabilities, highlighting the practical role of TriSig to assess and guide the ongoing triclustering searches; and (3) generate synthetic data to assess what makes a tricluster statistically significant and how multi-hypotheses corrections affect outcomes.

Although our solution covers all the statistical principles needed for evaluating patterns in tensor data exhibiting constant and additive coherence, it does not cover principles to statistically assess patterns with order-preserving factors. While our methodology offers expressive default behavior, we recommend fine-tuning the underlying assumptions based on the discussed guidelines to enhance overall robustness.

Our work provides unique opportunities that go beyond the minimization of false positive discoveries. State-of-the-art triclustering algorithms can use the proposed set of statistical principles in their search criteria with the aim of focusing on statistically significant subspaces and thus aiding in pattern-centric descriptive and predictive tasks. The implementation of the proposed statistical tests can be found at https://github.com/JupitersMight/TriSig.

Table 5
Comparison of three statistical methodologies – TriSig, SimSig [23], and MoiseSig [22] – for the triclustering solutions previously produced in Fig. 2 based on the distribution of statistically significant patterns – p-value quartiles (Q1, median, Q3), reference significance threshold (critical p-value), and fraction of statistically significant triclusters. Note that SimSig and MoiseSig are applied over the normalized real-valued tensor data.

			Q1	Median	Q3	Critical p-value	Significant
δ-Trimax	*penicillin batch fermentation*	MoiseSig	1.61×10^{-175}	8.48×10^{-151}	1.44×10^{-125}	0.05	100%
		SimSig	1.31×10^{-4}	1.45×10^{-2}	1.30×10^{-1}	1.00×10^{-4}	18%
		TriSig	5.24×10^{-275}	7.16×10^{-103}	1.32×10^{-23}	2.71×10^{-19}	91%
	glycine	MoiseSig	2.23×10^{-308}	2.23×10^{-308}	2.23×10^{-308}	0.05	100%
		SimSig	4.91×10^{-1}	5.33×10^{-1}	5.41×10^{-1}	1.00×10^{-4}	0%
		TriSig	2.65×10^{-272}	8.14×10^{-92}	3.37×10^{-69}	3.36×10^{-69}	100%
	gene expression	MoiseSig	1.00	1.00	1.00	0.05	0%
		SimSig	≈ 0	≈ 0	≈ 0	1.00×10^{-04}	100%
		TriSig	2.85×10^{-27}	3.96×10^{-15}	1.00	2.08×10^{-9}	70%
TriCluster	*penicillin batch fermentation*	MoiseSig	2.23×10^{-308}	2.23×10^{-308}	2.23×10^{-308}	0.05	100%
		SimSig	1.83×10^{-1}	2.19×10^{-1}	2.46×10^{-1}	1.00×10^{-4}	0%
		TriSig	2.21×10^{-20}	6.05×10^{-18}	4.16×10^{-17}	4.26×10^{-12}	100%
	glycine	MoiseSig	6.90×10^{-12}	1.09×10^{-8}	3.13×10^{-6}	0.05	100%
		SimSig	5.51×10^{-2}	6.90×10^{-2}	9.16×10^{-2}	1.00×10^{-4}	0%
		TriSig	7.41×10^{-193}	1.63×10^{-183}	5.67×10^{-179}	3.28×10^{-176}	100%
	gene expression	MoiseSig	1.00	1.00	1.00	0.05	0%
		SimSig	≈ 0	≈ 0	≈ 0	1.00×10^{-04}	100%
		TriSig	1.00	1.00	1.00	0.05	0%
TRIAS	*penicillin batch fermentation*	MoiseSig	2.23×10^{-308}	7.14×10^{-140}	1.81×10^{-91}	0.05	100%
		SimSig	2.25×10^{-9}	5.89×10^{-6}	4.59×10^{-2}	1.00×10^{-4}	60%
		TriSig	1.00	1.00	1.00	0.003	22%
	glycine	MoiseSig	$2.23\times10^{-}308$	9.82×10^{-16}	6.90×10^{-12}	0.05	100%
		SimSig	4.33×10^{-1}	4.52×10^{-1}	4.98×10^{-1}	1.00×10^{-4}	0%
		TriSig	4.16×10^{-178}	2.87×10^{-173}	4.83×10^{-170}	2.05×10^{-168}	100%
	gene expression	MoiseSig	1.00	1.00	1.00	0.05	0%
		SimSig	≈ 0	≈ 0	≈ 0	1.00×10^{-4}	100%
		TriSig	1.00	1.00	1.00	0.05	0%

Future work will focus on: (1) extending these principles to N-way tensor data, bridging the current gap on how to assess the statistical significance of patterns in tensor data with more than three dimensions, and (2) evaluating the impact on pattern formation when this methodology is incorporated into the search criteria of triclustering algorithms.

Funding

This work was supported by Fundação para a Ciência e a Tecnologia (FCT) under the PhD grant to LA (2021.07759.BD), INESC-ID pluriannual (UIDB/50021/2020), and Associate Laboratory for Green Chemistry (LAQV) financed by national funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020), and the contract CEECIND/01399/2017/CP1462/CT0015 to RSC. The authors also wish to acknowledge the European Union's Horizon BioLaMer project under grant agreement 101099487.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Availability

The code is freely available at https://github.com/JupitersMight/TriSig under the MIT license.

Appendix

See Table A.6.

Table A.6
Illustrative null model assumptions for diverse three-way data domains. Data is described according to *observations-variables-context* axes and the underlying satisfied assumptions.

Data	Variables MD Context MD	Variables MD Context MI	Variables MI Context MD	Variables MD Context TC	Variables MI Context TC	Variables MI Context MI
station-day cluster-month [46]	✓	✓				✓
movie-keywords-celeb [47]	✓	✓				✓
gene-marker-tumor [48,49]	✓	✓				✓
subject-fMRI connectivity pair-time window [15]	✓	✓				✓
job-jobseeker-seeker skill [47]		✓				✓
individual-signal-signal [50]		✓				✓
position-position-variable [11,12]		✓				✓
movie-genre-tag [51]		✓				✓
user-behavior-environment [52]		✓				✓
stations/provider-variables-time [10]				✓	✓	
sample-gene-time [5,6,53–55]				✓	✓	
patient/provider-variable-time [9,14]				✓	✓	
vertex-vertex-time [21]					✓	
users-bookmarks-tags (topics) [8,47]						✓

References

[1] R. Henriques, S.C. Madeira, Triclustering algorithms for three-dimensional data analysis: a comprehensive survey, ACM Comput. Surv. 51 (5) (2018) 1–43.

[2] D.F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, S.C. Madeira, Learning prognostic models using a mixture of biclustering and triclustering: Predicting the need for non-invasive ventilation in amyotrophic lateral sclerosis, J. Biomed. Inform. 134 (2022) 104172.

[3] D.F. Soares, R. Henriques, M. Gromicho, M. de Carvalho, S.C. Madeira, Triclustering-based classification of longitudinal data for prognostic prediction: targeting relevant clinical endpoints in amyotrophic lateral sclerosis, Sci. Rep. 13 (1) (2023) 6182.

[4] J.M. White, M.J. Piron, V.R. Rangaraj, E.C. Hanlon, R.N. Cohen, M.J. Brady, Reference gene optimization for circadian gene expression analysis in human adipose tissue, J. Biol. Rhythms 35 (1) (2020) 84–97.

[5] S.C. Groen, I. Ćalić, Z. Joly-Lopez, A.E. Platts, J.Y. Choi, M. Natividad, K. Dorph, W.M. Mauck, B. Bracken, C.L.U. Cabral, et al., The strength and pattern of natural selection on gene expression in rice, Nature 578 (7796) (2020) 572–576.

[6] J. Liu, M. Frochaux, V. Gardeux, B. Deplancke, M. Robinson-Rechavi, Inter-embryo gene expression variability recapitulates the hourglass pattern of evo-devo, BMC Biol. 18 (1) (2020) 1–12.

[7] M. Yalçin, R. El-Athman, K. Ouk, J. Priller, A. Relógio, Analysis of the circadian regulation of cancer hallmarks by a cross-platform study of colorectal cancer time-series data reveals an association with genes involved in Huntington's disease, Cancers 12 (4) (2020) 963.

[8] D. Gnatyshak, D.I. Ignatov, A. Semenov, J. Poelmans, Gaining insight in social networks with biclustering and triclustering, in: International Conference on Business Informatics Research, Springer, 2012, pp. 162–171.

[9] T. Song, Q. Tang, J. Huang, Triadic closure, homophily, and reciprocation: an empirical investigation of social ties between content providers, Inf. Syst. Res. 30 (3) (2019) 912–926.

[10] M.H. Kazemi, A. Majnooni-Heris, O. Kisi, J. Shiri, Generalized gene expression programming models for estimating reference evapotranspiration through cross-station assessment and exogenous data supply, Environ. Sci. Pollut. Res. 28 (6) (2021) 6520–6532.

[11] J.L. Amaro-Mellado, L. Melgar-García, C. Rubio-Escudero, D. Gutiérrez-Avilés, Generating a seismogenic source zone model for the Pyrenees: A GIS-assisted triclustering approach, Comput. Geosci. 150 (2021) 104736.

[12] L. Melgar-García, D. Gutiérrez-Avilés, M.T. Godinho, R. Espada, I.S. Brito, F. Martínez-Álvarez, A. Troncoso, C. Rubio-Escudero, A new big data triclustering approach for extracting three-dimensional patterns in precision agriculture, Neurocomputing (2022).

[13] L. Alexandre, R.S. Costa, L.L. Santos, R. Henriques, Mining pre-surgical patterns able to discriminate post-surgical outcomes in the oncological domain, IEEE J. Biomed. Health Inf. 25 (7) (2021) 2421–2434.

[14] D. Soares, R. Henriques, M. Gromicho, S. Pinto, M.d. Carvalho, S.C. Madeira, Towards triclustering-based classification of three-way clinical data: A case study on predicting non-invasive ventilation in als, in: International Conference on Practical Applications of Computational Biology & Bioinformatics, Springer, 2020, pp. 112–122.

[15] M.A. Rahaman, E. Damaraju, J.A. Turner, T.G. van Erp, D. Mathalon, J. Vaidya, B. Muller, G. Pearlson, V.D. Calhoun, Tri-clustering dynamic functional network connectivity identifies significant schizophrenia effects across multiple states in distinct subgroups of individuals, Brain Connect. 12 (1) (2022) 61–73.

[16] G. Ciaburro, G. Iannace, Machine learning-based algorithms to knowledge extraction from time series data: A review, Data 6 (6) (2021) 55.

[17] K. Sim, G.E. Yap, D.R. Hardoon, V. Gopalkrishnan, G. Cong, S. Lukman, Centroid-based actionable 3D subspace clustering, IEEE Trans. Knowl. Data Eng. 25 (6) (2012) 1213–1226.

[18] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature Genet. 25 (1) (2000) 25–29.

[19] A.B. Tchagang, S. Phan, F. Famili, H. Shearer, P. Fobert, Y. Huang, J. Zou, D. Huang, A. Cutler, Z. Liu, et al., Mining biological information from 3D short time-series gene expression data: the OPTricluster algorithm, BMC Bioinformatics 13 (1) (2012) 1–17.

[20] D. Amar, D. Yekutieli, A. Maron-Katz, T. Hendler, R. Shamir, A hierarchical Bayesian model for flexible module discovery in three-way time-series data, Bioinformatics 31 (12) (2015) i17–i26.

[21] R. Guigoures, M. Boullé, F. Rossi, Discovering patterns in time-varying graphs: a triclustering approach, Adv. Data Anal. Classif. 12 (2018) 509–536.

[22] G. Moise, J. Sander, Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering, in: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2008, pp. 533–541.

[23] K. Sim, Z. Aung, V. Gopalkrishnan, Discovering correlated subspace clusters in 3D continuous-valued data, in: 2010 IEEE International Conference on Data Mining, IEEE, 2010, pp. 471–480.

[24] S. Mankad, G. Michailidis, Biclustering three-dimensional data arrays with plaid models, J. Comput. Graph. Statist. 23 (4) (2014) 943–965.

[25] D. Gutiérrez-Avilés, C. Rubio-Escudero, LSL: A new measure to evaluate triclusters, in: 2014 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2014, pp. 30–37.

[26] D. Gutiérrez-Avilés, C. Rubio-Escudero, MSL: a measure to evaluate three-dimensional patterns in gene expression data, Evol. Bioinform. 11 (2015) EBO–S25822.

[27] B.S. Biswal, S. Patra, A. Mohapatra, S. Vipsita, Trirnsc: triclustering of gene expression microarray data using restricted neighbourhood search, IET Syst. Biol. 14 (6) (2020) 323–333.

[28] Y. Wang, W. Pang, Z. Jiao, An adaptive mutual K-nearest neighbors clustering algorithm based on maximizing mutual information, Pattern Recognit. 137 (2023) 109273.

[29] D. Paul, S. Saha, J. Mathew, Fusion of evolvable genome structure and multi-objective optimization for subspace clustering, Pattern Recognit. 95 (2019) 58–71.

[30] S.C. Madeira, A.L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans. Comput. Biol. Bioinform. 1 (1) (2004) 24–45.

[31] R. Henriques, S.C. Madeira, BSig: evaluating the statistical significance of biclustering solutions, Data Min. Knowl. Discov. 32 (1) (2018) 124–161.

[32] R. Henriques, F.L. Ferreira, S.C. Madeira, BicPAMS: software for biological data analysis with pattern-based biclustering, BMC Bioinformatics 18 (1) (2017) 1–16.

[33] S.C. Madeira, M.C. Teixeira, I. Sa-Correia, A.L. Oliveira, Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm, IEEE/ACM Trans. Comput. Biol. Bioinform. 7 (1) (2008) 153–165.

[34] J.P. Gonçalves, S.C. Madeira, e-bimotif: Combining sequence alignment and biclustering to unravel structured motifs, in: Advances in Bioinformatics: 4th International Workshop on Practical Applications of Computational Biology and Bioinformatics 2010, IWPACBB 2010, Springer, 2010, pp. 181–191.

[35] G. Chetty, M. Chetty, Multiclass microarray gene expression analysis based on mutual dependency models, in: IAPR International Conference on Pattern Recognition in Bioinformatics, Springer, 2009, pp. 46–55.

[36] M.E. Mann, E.A. Lloyd, N. Oreskes, Assessing climate change impacts on extreme weather events: the case for an alternative (Bayesian) approach, Clim. Chang. 144 (2017) 131–142.

[37] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. Ser. B Methodol. 57 (1) (1995) 289–300.

[38] A. Bhar, M. Haubrock, A. Mukhopadhyay, E. Wingender, Multiobjective triclustering of time-series transcriptome data reveals key genes of biological processes, BMC Bioinformatics 16 (1) (2015) 1–19.

[39] R. Jaschke, A. Hotho, C. Schmitz, B. Ganter, G. Stumme, Trias–an algorithm for mining iceberg tri-lattices, in: Sixth International Conference on Data Mining, ICDM'06, IEEE, 2006, pp. 907–911.

[40] D. Soares, R. Henriques, M. Gromicho, S. Pinto, M. de Carvalho, S.C. Madeira, Towards triclustering-based classification of three-way clinical data: A case study on predicting non-invasive ventilation in als, in: Practical Applications of Computational Biology & Bioinformatics, 14th International Conference (PACBB 2020) 14, Springer, 2021, pp. 112–122.

[41] C. Irwin, M. Van Reenen, S. Mason, L.J. Mienie, J.A. Westerhuis, C.J. Reinecke, Contribution towards a metabolite profile of the detoxification of benzoic acid through glycine conjugation: an intervention study, PLoS One 11 (12) (2016) e0167309.

[42] J. Kanno, K.-i. Aisaki, K. Igarashi, N. Nakatsu, A. Ono, Y. Kodama, T. Nagao, "Per cell" normalization method for mRNA measurement by quantitative PCR and microarrays, BMC Genomics 7 (1) (2006) 1–14.

[43] S. Goldrick, A. Ştefan, D. Lovett, G. Montague, B. Lennox, The development of an industrial-scale fed-batch fermentation simulation, J. Biotechnol. 193 (2015) 70–82.

[44] S. Goldrick, C.A. Duran-Villalobos, K. Jankauskas, D. Lovett, S.S. Farid, B. Lennox, Modern day monitoring and control challenges outlined on an industrial-scale benchmark fermentation process, Comput. Chem. Eng. 130 (2019) 106471.

[45] J. Lobo, R. Henriques, S.C. Madeira, G-Tric: generating three-way synthetic datasets with triclustering solutions, BMC Bioinformatics 22 (1) (2021) 1–28.

[46] X. Wu, C. Cheng, R. Zurita-Milla, C. Song, An overview of clustering methods for geo-referenced time series: From one-way clustering to co-and tri-clustering, Int. J. Geogr. Inf. Sci. 34 (9) (2020) 1822–1848.

[47] D.I. Ignatov, D.V. Gnatyshak, S.O. Kuznetsov, B.G. Mirkin, Triadic formal concept analysis and triclustering: searching for optimal patterns, Mach. Learn. 101 (1) (2015) 271–302.

[48] Y. Gan, Z. Dong, X. Zhang, G. Zou, Tri-clustering analysis for dissecting epigenetic patterns across multiple cancer types, in: International Conference on Intelligent Computing, Springer, 2018, pp. 330–336.

[49] M. Cardoso-Moreira, J. Halbert, D. Valloton, B. Velten, C. Chen, Y. Shao, A. Liechti, K. Ascenção, C. Rummel, S. Ovchinnikova, et al., Gene expression across mammalian organ development, Nature 571 (7766) (2019) 505–509.

[50] H. Joo, T. Simon, M. Cikara, Y. Sheikh, Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10873–10883.

[51] D.V. Gnatyshak, Greedy modifications of oac-triclustering algorithm, Procedia Comput. Sci. 31 (2014) 1116–1123.

[52] S.J. Ahn, J. Kim, J. Kim, The bifold triadic relationships framework: A theoretical primer for advertising research in the metaverse, J. Advert. 51 (5) (2022) 592–607.

[53] B. Strober, R. Elorbany, K. Rhodes, N. Krishnan, K. Tayeb, A. Battle, Y. Gilad, Dynamic genetic regulation of gene expression during cellular differentiation, Science 364 (6447) (2019) 1287–1290.
[54] J.K. Kim, Y. Chen, A.J. Hirning, R.N. Alnahhas, K. Josić, M.R. Bennett, Long-range temporal coordination of gene expression in synthetic microbial consortia, Nat. Chem. Biol. 15 (11) (2019) 1102–1109.
[55] K. Mandal, R. Sarmah, D.K. Bhattacharyya, POPTric: Pathway-based order Preserving Triclustering for gene sample time data analysis, Expert Syst. Appl. 192 (2022) 116336.

Leonardo Alexandre received a B.Sc. degree in computer science and engineering from the Instituto Superior de Engenharia de Lisboa (ISEL), Lisboa, Portugal, in 2018, and the M.Sc. degree in computer science and engineering at Instituto Superior Técnico (IST) Universidade de Lisboa, Lisboa, Portugal, in 2021. He is currently developing his Ph.D. in computer science and engineering, with a focus on Pattern Discovery.

R. Costa is currently an assistant researcher at FCT-NOVA. He also holds a lecturer position. His research focuses on systems biology modeling and data science with application in biotechnology, (bio)process engineering and healthcare.

Rui Henriques is a Professor at the Instituto Superior Técnico, University of Lisbon, where he lectures courses in the domains of Machine Learning, Artificial Intelligence, and Health Systems. Graduated (2008) and doctorate (2016) by Instituto Superior Técnico, Rui had wide-exposure to real projects in different sectors as an analyst at McKinsey and head of the AI area at Tekever. Rui is also an Integrated Researcher at INESC-ID, where he coordinates multidisciplinary R&D projects. Rui has been ascribed scientific career awards in 2020 and 2022. His research contributions cross advances in machine learning with real-world applications in biomedical and urban domains.