ACCSCIENCE PUBLISHING

Artificial Intelligence in Health

ORIGINAL RESEARCH ARTICLE

Integrated sources model: A new space-learning model for heterogeneous multi-view data reduction, visualization, and clustering

Supplementary File

1. A variant of the Integrated Sources Model to handle missing views

An important limitation of the Integrated Sources Model (ISM) and other multi-view latent space approaches is the requirement for the availability of multi-view data for all observations in the training dataset. For financial and/or logistical reasons, a particular view may be missing in a subset of the observations, and this subset may vary depending on the view under consideration.

Workflow S1 describes a variant of ISM that can process multi-view data with missing views. In this approach, ISM itself is applied to a collection of ISM-transformed datasets, each derived from a subset of views whose intersection contains at least a suitable number of observations to be processed by ISM. Within each subset, an additional expansion process allows the integration of all observations inside and outside each view, resulting in much larger transformed views than the original intersection would allow, as shown in Figure S1.

Workflow S1. Variant of ISM to handle missing views

Input: m views $\{X_1, \cdots, X_m\}, X^v \in \mathbb{R}_+^{n \times d_v}$ where n is the number of rows common to all views and d_v is the number of columns in the v^{th} view (it is assumed for each column that its values lie between 0 and 1 after normalization by the maximum row value).

Output: NTF factors $W^*, H^*, Q^*, W^* \in \mathbb{R}_+^{n_+ \times d_l}, H^* \in \mathbb{R}_+^{d_l \times d_l}, Q^* \in \mathbb{R}_+^{m \times d_l}$ and updated view-mapping matrix H where n_+ is the number of rows in the union of all observations in all views and d_l is the dimension of the latent space.

1: Partition: Create subsets of views, each with an intersection of views that contains a suitable number of observations to be processed by ISM;

2: Local integration: Apply ISM on each subset of views;

3: Projection: For each view in a given subset, project the non-missing observations outside the intersection onto the latent space by using the ISM Workflow 2;
(For other views for which the corresponding observations are missing, ISM view- scores remain missing)

4: Expansion: Estimate missing view scores by the weighted average of existing view scores, where the weights are the ISM view loadings;

5: Unified Integration: Apply ISM to the expanded transformed data from all subsets of views;

Abbreviations: ISM: Integrated Sources Model; NTF: Non-negative tensor factorization.

2. Application of the Integrated Sources Model variant to the UCI Digits data

UCI Digits data: The data can be found at https://archive.ics.uci.edu/datasets and contains six heterogeneous views: 76 Fourier coefficients of the character shapes, 216 profile correlations, 64 Karhunen-Love coefficients, 240-pixel averages of the images from 2 × 3 windows, 47 Zernike moments, and six morphological features. Each class contains 200 labeled examples.

In the first two views, the last 500 examples were set to missing. Two sets of views were considered, consisting of the first three and last three views, respectively. Two separate ISM analyses were performed for each view-set. In the first analysis, the first 1500 examples were included, while in the second analysis, the last 1500 examples were included in the study. Thus, only 1000 out of the 2000 examples were analyzed with all available views.

After applying the ISM expansion process, the two ISM-transformed data containing the 2000 examples were integrated using ISM to obtain the meta-scores. Following the article's analysis workflow, the 10 classes were identified (Figure S2) with a purity index of 5.31, which is slightly lower than the purity index of 5.81 obtained in the original data analysis.

Robustness results for multi-view multidimensional scaling, principal component analysis, group factor analysis, and Multi-Omics Factor Analysis+ with respect to the chosen rank.

To assess the robustness of the results with respect to the chosen rank, we tried a range of values for the rank around the elbow observed in the variance ratio scree plot.

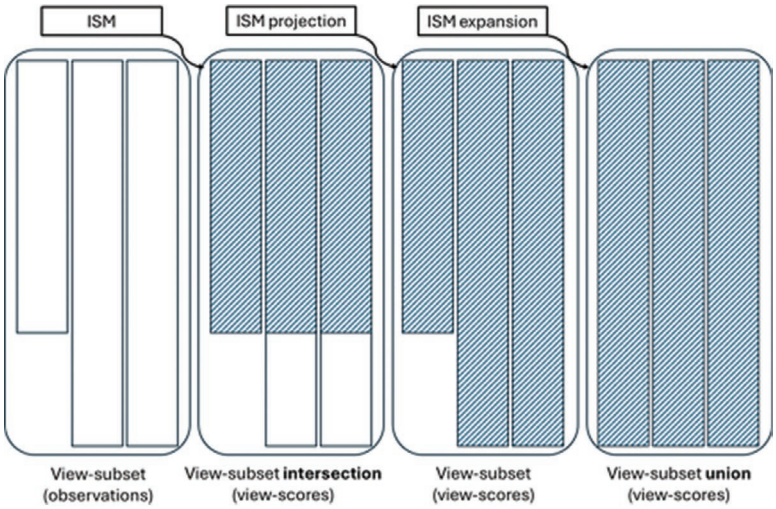

Figure S1. Illustration of the Integrated Sources Model (ISM) expansion process

Table S1. Performance metrics for four latent-space methods as a function of the rank used

Method	Rank	Class retrieval	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall performance
MVMDS	9	0.80	0.48	0.53	0.63	0.58	0.62	0.22	0.55
MVMDS	10	0.70	0.41	0.49	0.61	0.54	0.62	0.21	0.51
MVMDS	11	0.90	0.46	0.47	0.56	0.52	0.61	0.21	0.53
PCA	9	0.40	0.19	0.44	0.57	0.51	0.73	0.38	0.46
PCA	10	0.40	0.19	0.44	0.57	0.51	0.73	0.38	0.46
PCA	11	0.40	0.19	0.43	0.57	0.51	0.73	0.38	0.46
GFA	8	0.90	0.45	0.48	0.61	0.54	0.32	0.15	0.49
GFA	9	0.90	0.52	0.54	0.64	0.59	0.34	0.14	0.52
GFA	10	0.80	0.39	0.45	0.58	0.51	0.34	0.12	0.46
MOFA+	9	0.40	0.13	0.26	0.37	0.36	0.33	0.15	0.29
MOFA+	10	0.70	0.29	0.36	0.46	0.44	0.34	0.13	0.39
MOFA+	11	0.40	0.13	0.27	0.39	0.35	0.34	0.12	0.29

Abbreviations: ARI: Adjusted Rand index; GFA: Group factor analysis; FMS: Fowlkes-Mallows score; MOFA+: Multi-Omics Factor Analysis+; MVMDS: Multi-view multidimensional scaling; NMI: Normalized Mutual Information index; PCA: Principal component analysis.

Table S2. Performance metrics for four latent-space methods as a function of the rank used

Method	Rank	Class retrieval	Purity	ARI	NMI	FMS	Sparsity	Specificity	Overall performance
MVMDS	9	0.75	0.67	0.96	0.94	0.97	0.56	0.21	0.72
MVMDS	10	0.75	0.70	0.97	0.95	0.97	0.56	0.21	0.73
MVMDS	11	0.69	0.64	0.98	0.96	0.98	0.59	0.21	0.72
PCA	9	0.63	0.41	0.92	0.86	0.93	0.57	0.24	0.65
PCA	10	0.56	0.40	0.94	0.89	0.95	0.57	0.23	0.65
PCA	11	0.63	0.48	0.95	0.91	0.96	0.60	0.22	0.68
GFA	11	0.81	0.69	0.97	0.94	0.97	0.33	0.13	0.69
GFA	12	0.81	0.74	0.98	0.96	0.98	0.30	0.09	0.69
GFA	13	0.81	0.71	0.97	0.95	0.97	0.38	0.08	0.70
MOFA+	12	0.69	0.61	0.94	0.92	0.95	0.55	0.20	0.69
MOFA+	13	0.81	0.76	0.94	0.93	0.95	0.56	0.19	0.73
MOFA+	14	0.69	0.52	0.81	0.89	0.85	0.49	0.17	0.63

Abbreviations: ARI: Adjusted Rand index; GFA: Group factor analysis; FMS: Fowlkes-Mallows score; MOFA+: Multi-Omics Factor Analysis+; MVMDS: Multi-view multidimensional scaling; NMI: Normalized Mutual Information index; PCA: Principal component analysis.

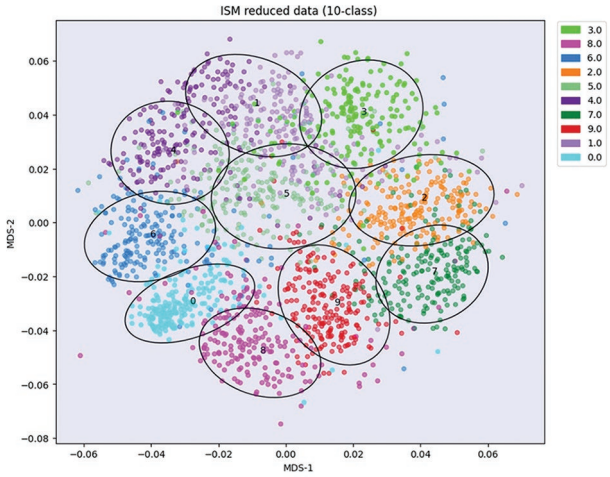

Figure S2. Analysis of the UCI Digits dataset using the ISM expansion process after masking a large number of views.
Abbreviations: ISM: Integrated Sources Model; MDS: Multidimensional scaling.

The performance measures are presented in Table S1 (UCI Digits data) and Table S2 (Signature 915 data).

3. Robustness results for multi-view multidimensional scaling, principal component analysis, group factor analysis, and Multi-Omics Factor Analysis+ (UCI Digits data, 10 classes)
4. Robustness results for multi-view multidimensional scaling, principal component analysis, group factor analysis, and Multi-Omics Factor Analysis+ (Signature 915 data, 16 classes)
5. Distributed non-negative matrix factorization: Workflow

Workflow S2. Distributed NMF using ISM

Input: matrix $X \in \mathbb{R}_+^{n \times p}$.

Output: factoring matrices $W \in \mathbb{R}_+^{n \times k}, H \in \mathbb{R}_+^{p \times k}$.

1: Consider a random partition of X into m matrices $X_v \in \mathbb{R}_+^{n \times p_v}$ (matrices may be of different size, e.g., p/m if is not an integer).

2: Factorize each view X_v using NMF and same rank k:

$X_v = W_v H_v^T + E_v, W_v \in \mathbb{R}_+^{n \times k}, H_v \in \mathbb{R}_+^{p_v \times k}, E_v \in \mathbb{R}^{n \times p_v}$.

3: Apply ISM on the list of views $\{W_v\}_{1 \le v \le m}$:

$W_v = W^* H_v^{*T} + E_v^*, W^* \in \mathbb{R}_+^{n \times k}, H_v^* \in \mathbb{R}_+^{k \times k}, E_v^* \in \mathbb{R}^{n \times k}$ where W^* contains ISM meta- scores $\left\{H_v^*\right\}_{1 \le v \le m}$ and contains the view-mapping matrices to the $\{W_v\}_{1 \le v \le m}$.

4: Factorize each view X_v using H_v from step 2 and $W^* H_v^{*T}$ from step 3:

$X_v = W^* H_v^{*T} H_v^T + E_v^{'}, W^* \in \mathbb{R}_+^{n \times k}, H_v^* \in \mathbb{R}_+^{k \times k}, H_v \in \mathbb{R}_+^{p_v \times k}, E_v^{'} \in \mathbb{R}^{n \times p_v}$.

5: X can now be factorized:

$X = WH^T + E, W = W^* \in \mathbb{R}_+^{n \times k}, H = \left\{H_v H_v^*\right\}_{1 \le v \le m} \in \mathbb{R}_+^{p \times k}, E = \left\{E_v^{'}\right\}_{1 \le v \le m} \in \mathbb{R}^{n \times p}$.

Abbreviations: ISM: Integrated Sources Model; NMF: Non-negative matrix factorization.

6. Distributed non-negative matrix factorization: Example

A dense matrix of size 76 × 10,000 was analyzed using either standard non-negative matrix factorization (NMF) or distributed NMF with 10 slices of size 76 × 1000. When four components were used, the relative errors were very similar (0.40 for NMF vs. 0.41 for distributed NMF, respectively). However, the computational time required by distributed NMF was reduced by 13% when separate factorizations were performed in a sequential way, and by 92% when separate factorizations were performed in parallel.