Factor Analysis with Correlated Topic Model for Multi-Modal Data

Małgorzata Łazęcka
Faculty of Mathematics, Informatics
and Mechanics,
University of Warsaw
Institute of Computer Science,
Polish Academy of Sciences
m.lazecka@uw.edu.pl

Ewa Szczurek
Institute of AI for Health,
Helmholtz Center Munich
Faculty of Mathematics, Informatics
and Mechanics,
University of Warsaw
em.szczurek@uw.edu.pl

Abstract

Integrating various data modalities brings valuable insights into underlying phenomena. Multimodal factor analysis (FA) uncovers shared axes of variation underlying different simple data modalities, where each sample is represented by a vector of features. However, FA is not suited for structured data modalities, such as text or single cell sequencing data, where multiple data points are measured per each sample and exhibit a clustering structure. To overcome this challenge, we introduce FACTM, a novel, multi-view and multi-structure Bayesian model that combines FA with correlated topic modeling and is optimized using variational inference. Additionally, we introduce a method for rotating latent factors to enhance interpretability with respect to binary features. On text and video benchmarks as well as real-world music and COVID-19 datasets, we demonstrate that FACTM outperforms other methods in identifying clusters in structured data, and integrating them with simple modalities via the inference of shared, interpretable factors.

1 INTRODUCTION

Real-world data often spans multiple modalities, each capturing distinct yet complementary aspects of the underlying phenomena. Analysis of such multi-modal data has emerged as a critical machine learning task across various application domains. For example, video data combine temporal image sequences, audio signals, and transcriptions of spoken content, while medical datasets integrate diverse patient measurements ranging from electronic health records to imaging (e.g., computed tomography, CT), and molecular data (e.g., single-cell RNA sequencing, scRNA-seq). Effective synthesis of these heterogeneous information sources is crucial for comprehensive sample characterization and predictive modeling.

A state of the art approach to interpretable data integration is Factor Analysis (FA) and its multi-view extensions, where each view corresponds to a modality. These methods perform unsupervised factorization of high-dimensional observations into interpretable latent factors and weights. FA serves multiple analytical objectives: it reduces data dimensionality, enables integration of heterogeneous data sources, and reveals principal axes of variation across samples, facilitating interpretation of complex datasets.

Despite these advantages, FA-based approaches face several fundamental limitations. First, FA models are restricted to handling *simple* data, where for each data view the samples are described by vectors. However, frequently, apart from simple views, multi-modal data views fall into a category of *structured* views, where each sample consists of a set of data points. These data points arise from an underlying cluster structure and are characterized by values assigned to observed objects. For such structured views, each sample can be summarized by a vector of cluster abundances. Such structured data arise in various applications, including text document analysis, where documents contain multiple sentences composed of words. Sentences cluster into topics, and documents are represented by topic distributions. Another example is scRNA-seq data, where samples contain multiple cells with gene expression values. Here, cells cluster into distinct cell

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

types, and samples are characterized by cell-type proportions. Despite the abundance of such real-world data, there exists no FA-based approach capable of unified modeling of simple and structured views. The second limitation of FA is lack of full identifiability in terms of latent factors, as they may be permuted, their signs may be switched, or the weight and factor matrices may be rotated without affecting the model likelihood.

To address these challenges, we introduce FACTM, a novel Bayesian method designed for joint modeling of both simple and structured data across multiple views. FACTM extends the multimodal FA by leveraging the Correlated Topic Model (CTM), originally developed for text mining, to identify clusters and their prevalences within structured views. To enable integration across structured and simple views, we link the FA and CTM parts of the model through dedicated variables that are interpreted as sample-specific modifications to the population-level cluster proportions. To address the rotation invariance issue in a way that enhances model identifiability, we propose a supervised rotation method that incorporates additional features with which the factors are expected to be associated.

Our contributions are as follows: (i) we propose a novel model capable of handling both simple and structured views; (ii) for structured views, our model infers a covariance matrix that reveals relationships between the identified clusters; (iii) we introduce a method for meaningful supervised rotation, that enhances interpretability of latent factors in FA models; and (iv) through simulations, benchmark datasets, and real-world data, we demonstrate that FACTM outperforms existing methods and we show its practical utility.

2 RELATED WORK

Factor analysis FA (Thurstone, 1931) is a statistical method of representing data through latent factors, with probabilistic PCA (Tipping and Bishop, 1999) constituting a well-known example. An extension of FA suitable for high-dimensional data is the Tucker decomposition, which factors a tensor into component tensors. However, the Tucker decomposition is constrained by the requirement that the number of features in each view must be equal, rendering it inapplicable when these dimensions differ.

Numerous generalizations of FA have been developed to handle multiple data modalities, provided as separate data views with differing numbers of features, among which Bayesian approaches have proven particularly successful. Notable examples include Group Factor Analysis (GFA) (Klami et al., 2014) and Multi-Omics Factor Analysis (MOFA) (Argelaguet et al., 2018). Both models use automatic relevance determination to enforce factor-wise sparsity (allowing some factors to be inactive across some views). Additionally, MOFA employs spike-and-slab prior to shrink individual loadings to zero. Recent advancements, such as BASS (Zhao et al., 2016) and MuVI (Qoku and Buettner, 2023), introduce structured sparsity assumptions or domain-informed priors for the weights, while MEFISTO (Velten et al., 2022) extends MOFA to account for spatio-temporal dependencies. The addition of priors limits FA non-identifiability as they impose constraints on the possible weight and factor matrices.

We emphasize that the introduction of sparsity-inducing priors is crucial as it limits FA non-identifiability by imposing constraints on the possible weight and factor matrices. Additionally, these priors serve as an effective tool for denoising the data. Apart from FA, priors enforcing sparsity were shown to enhance performance in deep neural networks and were successfully applied in recent Variational Autoencoder models (e.g. Tonolini et al. (2020); Fallah and Rozell (2022))

The optimization strategies of the Bayesian FA models are based on Gibbs sampling or variational inference, which ranges from analytically derived EM-like updates of the variational parameters to automated variational inference methods.

Topic models Topic models are widely used methods of unsupervised learning of text documents representations, based on a clustering of words into topics. Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a standard topic model, assumes that the words in each document are drawn from a mixture of topics, which are shared across all documents and defined as distributions over words. The topic proportions are document-specific and are generated from a Dirichlet distribution. We highlight two extensions of LDA. The first, Correlation Topic Model (CTM) (Lafferty and Blei, 2005), uses more flexible distribution for the topic proportions to enable the inference of a covariance structure among the topics, allowing for the presence of one topic to be correlated with the presence of another. The second extension, ProdLDA (Srivastava and Sutton, 2017), replaces the LDA's mixture of topics with a weighted product of experts. Apart from text, topic models were applied also to molecular biological datasets, for example to identify tissue microenvironments (Chen et al., 2020) or deconvolve cell types from multi-cellular pixel resolution data in spatial transcriptomics (Miller et al., 2022).

3 DESCRIPTION OF METHODS

3.1 Background

Factor analysis A standard FA model aims to linearly reduce dimensionality of data while preserving the main axes of variation by factorizing a single given data matrix $Y \in \mathbb{R}^{N \times D}$, into two matrices: $Z \in \mathbb{R}^{N \times K}$ with latent factors, and $W \in \mathbb{R}^{D \times K}$ with weights (factor loadings), where N denotes the number of samples, D the number of features, and K the number of latent factors. This relationship can be expressed as $Y = ZW' + \varepsilon$, where ε captures random noise. We consider the data modality Y in standard FA as *simple*, since each sample n is described by a vector of D features, $Y_n \in \mathbb{R}^D$.

FA invariance Let $R \in \mathbb{R}^{K \times K}$ be a rotation matrix satisfying the property $R^{-1} = R'$. The likelihood in factor analysis is invariant under such rotations. After applying a rotation to both latent factors $\tilde{Z} = ZR$ and loadings $\tilde{W} = WR$, the likelihood remains unchanged, as

$$ZW' = ZRR'W' = \tilde{Z}\tilde{W}'.$$

The marginal distributions of Z and W are also invariant under isotropic normal priors. Moreover, even without assuming an isotropic prior on the loading matrix W, there is no guarantee that the variances are indeed unequal. Common approaches to address this issue are introducing sparsity constraints (e.g. Argelaguet et al. (2018); Qoku and Buettner (2023)) or applying specific rotations, such as the varimax rotation (Kaiser, 1958).

We note that the ordering and sign of latent factors is also non-identifiable. However, a change in sign does not affect the interpretation of the FA model, and the factors are usually sorted after the model is fitted by the total variance that they explain across all views.

Variational inference Variational inference proceeds by maximizing an evidence lower bound (ELBO) of the marginal log-likelihood $p(Y)$

$$\begin{aligned} ELBO(q) &:= \mathbb{E}_q \log p(X, Y) - \mathbb{E}_q \log q(X) \\ &= \log p(Y) - D_{KL}(q(X)||p(X|Y)) \leq \log p(Y), \end{aligned}$$

over a family of variational distributions $q(X)$ approximating $p(X|Y)$. Here, X and Y represent hidden and observed variables, respectively, and D_{KL} denotes Kullback–Leibler divergence. Under mean-field assumption the optimal variational distribution $\hat{q}_i$ that maximises the ELBO can be calculated as follows

$$\log \hat{q}(x_i) \propto \mathbb{E}_{-x_i} \log p(X, Y), \qquad (1)$$

where $\mathbb{E}_{-x_i}$ denotes the expected value with respect to the q distribution for all the variables X except for x_i. The standard approach is to use coordinate ascent, iteratively updating each variational parameter one at a time until ELBO convergence. For details see Blei et al. (2016) and Wainwright and Jordan (2008).

3.2 FACTM

FACTM (Fig. 1) extends FA for multiple simple and structured views. For simple data modalities, FACTM proceeds akin to other Bayesian multimodal FA models. Specifically, FACTM uses M matrices of observations $Y^m \in \mathbb{R}^{N \times D^m}$ (often referred to as views) instead of a single observation matrix Y used by the standard FA. The objective remains to identify common latent factors Z for all of the views, while also deriving view-specific loading matrices W^m

$$Y^m = Z(W^m)' + \varepsilon^m.$$

Following other Bayesian models, we choose the prior for Z_n as $\mathcal{N}(0, I)$, and assume ε_n is normally distributed with a diagonal covariance matrix. As in Argelaguet et al. (2018), in FACTM we assume factor- and feature-wise sparsity on loading matrices W^m using automatic relevance determination and spike-and-slab priors. Namely, each $w^m_{d,k}$ is modeled as the product $\tilde{w}^m_{d,k} \cdot s^m_{d,k}$, and the joint distribution of these two variables is given by

$$p(\tilde{w}^m_{d,k}, s^m_{d,k}) = \mathcal{N}(\tilde{w}^m_{d,k} | 0, 1/\alpha^m_k)\text{Ber}(s^m_{d,k} | \theta^m_k).$$

Additionally, we assume standard conjugate priors on the parameters in the equation above: $\alpha^m_k \sim \mathcal{G}(a^\alpha_0, b^\alpha_0)$ and $\theta^m_k \sim \text{Beta}(a^\theta_0, b^\theta_0)$, and similarly $\varepsilon^m_{n,d} \sim \mathcal{N}(0, 1/\tau^m_d)$, with a conjugate prior on τ^m_d, $\tau^m_d \sim \mathcal{G}(a^\tau_0, b^\tau_0)$ (all $a^\cdot_0, b^\cdot_0$ are hyperparameters).

In contrast to other multimodal FA models, FACTM accounts also for *structured* data views, where each sample n is represented by a set of I_n data points. Each data point i consists of J_i measured objects $\overline{y}_{n,i,j}$, each with an assigned type g, thus it can be represented by a set of count values $\{\overline{y}_{n,i,g}\}^G_{g=1}$, where $\overline{y}_{n,i,g} = \sum^{J_i}_j \mathbb{I}(\overline{y}_{n,i,j} = g)$. Extending the CTM, our model assumes that for each structured view, every sample (document in the CTM nomenclature) can be represented by a vector of abundances of L clusters (topics). This is modeled by the variable η_n, which is transformed via the softmax function to yield a probability distribution over the clusters: $\text{softmax}(\eta_n) = (\exp(\eta_{n,1})/C, \exp(\eta_{n,2})/C, \ldots, \exp(\eta_{n,L})/C)$, and $C = \sum^L_{l=1} \exp(\eta_{n,l})$. The variable η_n is drawn from a multivariate normal distribution $\mathcal{N}_L(\mu_n + \mu^{(0)}, \Sigma^{(0)})$, with $\Sigma^{(0)}$ accounting for the population-level covariance among the clusters. In contrast to standard CTM,

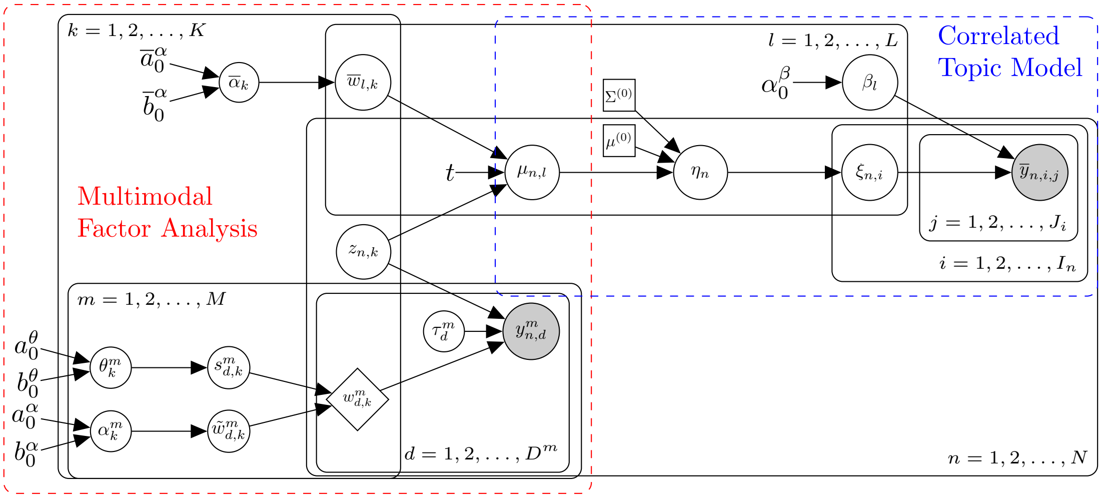

Figure 1: Graphical representation of FACTM. A single structured view is shown (in blue), although any number is possible.

we assume that the mean consists of two components: a sample-specific term μ_n, which modifies the population-level variable $\mu^{(0)}$ based on individual characteristics of observation n. This sample-specific term constitutes the link between the structured view and the remaining views in the model and depends on the factors Z and view-specific loadings $\overline{W}$ (Fig. 1). Given η_n, we sample the cluster assignment $\xi_{n,i}$ for each data point (sentence) i from a multinomial distribution: $\text{Mult}(1, \text{softmax}(\eta_n))$. Each cluster is characterized by a distribution over object types (distinct words), denoted by $g = 1, 2, \ldots, G$. These distributions (topic-word distributions) β_l are drawn from a Dirichlet distribution $\text{Dir}(\alpha)$. For each observed object (word) $\overline{y}_{n,i,j}$, knowing its cluster assignment $\xi_{n,i}$, we sample its type using the corresponding topic distribution $\text{Mult}(1, \beta_{\xi_{n,i}})$.

The notation used to describe FACTM is summarized in Table D.3 in the Appendix.

3.3 Inference

The joint probability distribution defining our model (Fig. 1) is given by

$$p(Z, W, Y, \overline{W}, \mu, \eta, \xi, \beta, \overline{Y}, \mathcal{X} | \mu^{(0)}, \Sigma^{(0)}, \mathcal{H}) =$$

$$\prod_{n=1}^{N} \prod_{k=1}^{K} \mathcal{N}(z_{n,k} | 0, 1) \tag{2a}$$

$$\prod_{m=1}^{M} \prod_{d=1}^{D_m} \prod_{k=1}^{K} \mathcal{N}(\tilde{w}_{d,k}^m | 0, 1/\alpha_k^m) \text{Ber}(s_{d,k}^m | \theta_k^m) \tag{2b}$$

$$\prod_{m=1}^{M} \prod_{k=1}^{K} \mathcal{G}(\alpha_k^m | a_0^\alpha, b_0^\alpha) \text{Beta}(\theta_k^m | a_0^\theta, b_0^\theta) \tag{2c}$$

$$\prod_{m=1}^{M} \prod_{n=1}^{N} \prod_{d=1}^{D_m} \mathcal{N}(y_{n,d}^m | \sum_{k=1}^{K} z_{n,k} w_{d,k}, 1/\tau_d^m) \tag{2d}$$

$$\prod_{m=1}^{M} \prod_{d=1}^{D_m} \mathcal{G}(\tau_d^m | a_0^\tau, b_0^\tau) \tag{2e}$$

$$\prod_{k=1}^{K} \prod_{l=1}^{L} \mathcal{N}(\overline{w}_{l,k} | 0, 1/\overline{\alpha}_k) \prod_{k=1}^{K} \mathcal{G}(\overline{\alpha}_k | \overline{a}_0^\alpha, \overline{b}_0^\alpha) \tag{2f}$$

$$\prod_{n=1}^{N} \prod_{l=1}^{L} \mathcal{N}(\mu_{n,l} | \sum_{k=1}^{K} z_{n,k} \overline{w}_{l,k}, 1/t) \tag{2g}$$

$$\prod_{n=1}^{N} \mathcal{N}_L(\eta_n | \mu_n + \mu^{(0)}, \Sigma^{(0)}) \tag{2h}$$

$$\prod_{n=1}^{N} \prod_{i=1}^{I_n} \text{Mult}(\xi_{n,i} | 1, \text{softmax}(\eta_n)) \tag{2i}$$

$$\prod_{n=1}^{N} \prod_{i=1}^{I_n} \prod_{j=1}^{J_i} \text{Mult}(\overline{y}_{n,i,j} | 1, \beta_{\xi_{n,i}}) \prod_{l=1}^{L} \text{Dir}(\beta_l | \alpha_0^\beta), \tag{2j}$$

where $\mathcal{X}$ denotes all the remaining nodes in $\{\alpha, \theta, \tau, \overline{\alpha}\}$, and $\mathcal{H}$ all the hyperparameters in $\{a_0^\cdot, b_0^\cdot, \overline{a}_0^\alpha, \overline{b}_0^\alpha, t\}$. By default we fix $t = 1$, $a_0^\alpha = b_0^\alpha = \overline{a}_0^\alpha = \overline{b}_0^\alpha = 1\text{e}{-3}$, $a_0^\tau = b_0^\tau = 1\text{e}{-3}$, $a_0^\theta = b_0^\theta = 1$, $\alpha_0^\beta = 1$. Note that the hyperparameters were selected to get non-informative priors, e.g. for the beta distribution, the parameters were chosen to make the distribution uniform.

We apply the following mean-field assumption on variational distribution q,

$$q(Z, W, \overline{W}, \mu, \eta, \xi, \beta, \mathcal{X}) = \prod_{n=1}^{N} \prod_{k=1}^{K} q(z_{n,k}) \tag{3}$$

$$\prod_{m=1}^{M}\prod_{k=1}^{K}\prod_{d=1}^{D_m} q(\tilde{w}^m_{d,k}, s^m_{d,k}) \prod_{m=1}^{M}\prod_{k=1}^{K} q(\alpha^m_k) q(\theta^m_k)$$

$$\prod_{m=1}^{M}\prod_{d=1}^{D_m} q(\tau^m_d) \prod_{k=1}^{K}\prod_{l=1}^{L} q(\overline{w}_{l,k}) \prod_{k=1}^{K} q(\overline{\alpha}_k)$$

$$\prod_{n=1}^{N} q_L(\mu_n) \prod_{n=1}^{N}\prod_{l=1}^{L} q(\eta_{n,l}) \prod_{n=1}^{N}\prod_{i=1}^{I_n} q(\xi_{n,i}) \prod_{l=1}^{L} q(\beta_l).$$

Additionally, we assume, that $q_L(\eta_n)$ follows normal distribution, and $q(\xi_{n,i})$ multinomial distribution. For the update equations of parameters of variational distributions from lines (2a)-(2f) we refer to Argelaguet et al. (2018), and for those in (2h)-(2j), we refer to Lafferty and Blei (2005); Blei and Lafferty (2007). Additionally, updates for $\mu^{(0)}$ and $\Sigma^{(0)}$ can be found in Masada and Takasu (2013). Below, we provide update equations for μ_n from (2g) for $n \in \{1, 2, \dots, N\}$. Note that the distribution $q_L(\mu_n)$ is multivariate. By applying (1), we have

$$\log(q_L(\mu_n)) \propto$$

$$\mathbb{E}_{-\mu_n} \log \Big(\mathcal{N}_L(\mu_n | \sum_{k=1}^{K} z_{n,k}\overline{w}_{\cdot,k}, \mathrm{diag}(1/t))$$

$$\cdot \mathcal{N}_L(\eta_n | \mu_n + \mu^{(0)}, \Sigma^{(0)}) \Big)$$

$$= \mathbb{E}_{-\mu_n} \log \mathcal{N}_L(\mu_n | \tilde{\mu}, \tilde{\Sigma})$$

with the parameters $\tilde{\Sigma}^{-1} = \mathrm{diag}(t) + \left(\Sigma^{(0)}\right)^{-1}$ and

$$\tilde{\mu} = \tilde{\Sigma} \left(\mathrm{diag}(t) \sum_{k=1}^{K} z_{n,k}\overline{w}_{\cdot,k} + \left(\Sigma^{(0)}\right)^{-1} (\eta_n - \mu^{(0)}) \right),$$

which follows from a standard computation for conjugate normal prior on the mean parameter and normal likelihood. Thus, after applying expected value, we get that the updates of the parameters of $q(\mu_n)$ are $\hat{\Sigma}^{(\mu)}_n = \left(\mathrm{diag}(t) + \left(\Sigma^{(0)}\right)^{-1}\right)^{-1}$,

$$\hat{\mu}^{(\mu)}_n = \hat{\Sigma}^{(\mu)}_n \Big(\mathrm{diag}(t) \sum_{k=1}^{K} \langle z_{n,k}\overline{w}_{\cdot,k} \rangle$$

$$+ \left(\Sigma^{(0)}\right)^{-1} (\langle \eta_n \rangle - \mu^{(0)}) \Big),$$

where $\langle \cdot \rangle$ denotes expected value with respect to q.

The computational complexity of FACTM inference is comparable to the individual components of the model (FA and CTM), as the parameters of the link variable μ_n are updated in a computationally efficient way, as shown in Figure B.1 in the Appendix.

3.3.1 Rotations

We propose a heuristic method for rotating the factors in order to increase their interpretability by associating them with a set of sample features. We solve this problem using the Kabsch-Umeyama algorithm (Kabsch, 1976). First we compute cross-correlation matrix $H \in \mathbb{R}^{K \times K}$ between K latent factors and K given features. Next, we perform an SVD decomposition on H, resulting in $H = USV'$. The rotation matrix is then given by $R = UV'$.

For numerical features, we use Pearson correlation r to compute each element of a matrix H. In case of binary features, we also use Pearson correlation r_{pb}, specifically known as the point-biserial correlation coefficient, which applies when one of the variables is binary, and the second one continuous. We motivate this choice by the property of r_{pb} that, after a monotonic transformation $r_{pb}((n_0 + n_1 - 2)/(1 - r^2_{pb}))^{1/2}$, where n_0 denotes the number of 0s and n_1 the number of 1s in the binary feature, it becomes a test statistic in the unpaired Student's t-test, comparing the means of the two groups.

4 EXPERIMENTS

We evaluated FACTM on comprehensive simulations, benchmark datasets and real-world data, measuring performance by: (i) estimation accuracy in versatile scenarios, (ii) learning data representations that are predictive in downstream classification tasks, (iii) efficacy of the rotation method in obtaining highly interpretable factors, (iv) ability to identify meaningful clusters (topics) in the structured data.

We provide code for the experiments and an implementation of FACTM on GitHub[1].

Compared methods FACTM was compared with several models, which either perform FA on simple views or topic modeling for structured data. For the FA, we used the following models: **FA Oracle**, serving as the upper bound for FA performance, which replaces the structured views by fixed simple views and fits a standard multimodal FA model. Specifically, for each structured view, we fix the μ_n variables to their true values (and not infer them as in FACTM) and represent the structured data as a simple view composed of the fixed μ_n vectors for the samples $n = 1, \dots, N$; **FA+CTM**, which proceeds in two steps: first, it fits CTM model to each structured view and represents the structured data as a simple view with the estimated $\eta_n - \mu^{(0)}$ as a feature vector for each sample n, and next applies FA to the original and such obtained simple views; and an ablation study that included **MOFA** (using the implementation from the package by Bredikhin et al. (2022)), **muVI** (with

[1] github.com/szczurek-lab/FACTM

the informed version of the model where applicable), **PCA** (Pedregosa et al., 2011), and **Tucker decomposition** (Kossaifi et al., 2019), which were fitted only to the original simple views (omitting the structured data). For the topic modeling, we compared FACTM with **CTM**, **LDA** (Pedregosa et al., 2011), and **ProdLDA** (implemented in the `pyro` package, Bingham et al. (2019)).

4.1 Parameter estimation accuracy

First, we evaluated the accuracy of parameter estimation using simulations.

Simulation settings The data were sampled from the model given by Equation (2), with the exception for factor-wise sparsity which was fixed. For the data generation, we used the following settings:: $N = 250$ samples, 3 views (consisting of 2 observed and 1 structured), each simple view with $D = 10$ features and each structured view with $L = 10$ topics. There were $K = 5$ true latent factors, sampled from a standard normal distribution, while the loading matrices were also sampled from standard normal with some columns set to zero (see Fig. B.2 in the Appendix) to introduce factor-wise sparsity. We applied low feature-wise sparsity, with 10% of the weights set to zero, resulting in up to 5 elements of W^m being zeroed out. The observations Y^m and the variable μ were sampled with a variance of 1. In the structured view, we used $G = 100$ distinct words, each sample (document) contained $I = 100$ data points (sentences) with $J = 10$ objects (words). The population-level variables were set to $\mu_l^{(0)} = 0$ for $l = 1, 2, \ldots, L$, and $\Sigma^{(0)}$ to 5 on the diagonal and 2.5 on the upper and lower diagonal, with all other elements set to 0. The topics were sampled from a Dirichlet distribution with equal parameters $\alpha = 1$. We devised the following simulation scenarios, each varying one parameter while keeping the rest fixed:

- Scenario 1: The link variable μ_n is multiplied by $\lambda \in \{0, 0.5, 1.5, 2\}$. Increasing λ strengthens the association of latent factors and structured views.
- Scenario 2: The Dirichlet distribution parameter for the topics, α, is varied within $\{5, 10\}$, with higher values leading to a more uniform topic distribution across words.
- Scenario 3: The number of topics is changed to $L \in \{5, 15\}$.
- Scenario 4: The variable $\mu^{(0)}$ is multiplied by $\lambda_{\mu^{(0)}}$, for $\lambda_{\mu^{(0)}} \in \{0.25, 0.5, 0.75, 1\}$. A higher $\lambda_{\mu^{(0)}}$ value results in greater disproportion in the baseline topic proportions. Here, $\mu^{(0)}$ is a centered, log-transformed vector scaled to sum to 1 of equally distributed values between 1 and 3.
- Scenario 5: The covariance matrix $\Sigma^{(0)}$ is scaled by $\lambda_{\Sigma^{(0)}} \in \{0.2, 0.6\}$.

The basic set of parameters is obtained by fixing $\lambda = 1$, $\alpha = 1$, $L = 10$, $\lambda_{\mu^{(0)}} = 0$, $\lambda_{\Sigma^{(0)}} = 1$.

Evaluation In the simulations, we first evaluated the inferred latent factors by comparing them to the ground truth. Given the permutation invariance of factors, we first computed Spearman correlation for all the estimated-true factor pairs and used the Hungarian method (Kuhn, 1955) to best match the inferred factors with the true ones (Fig. 2). We also evaluated the structured view part of the model by assessing how accurately it estimated the parameters of the latent variables compared to the true parameters, specifically focusing on the performance of the link variable μ_n (for the CTM model, we use $\eta_n - \mu^{(0)}$ as a substitute for the link variable μ_n, which is not explicitly estimated in that model), clustering variable ξ, topic distribution β (Fig. 3), and population-level variables $\mu^{(0)}$ and $\Sigma^{(0)}$ (Fig. 4). Since topic order is also not-identifiable, we applied the Hungarian method to the contingency table of true versus inferred topic assignments to determine the optimal ordering. Each simulation was repeated 10 times.

Results FACTM is more accurate in factor estimation compared to other models across all simulated scenarios (Fig. 2). For some settings, the additional information transferred from structured part modeled by CTM to FA in FA+CTM model actually deteriorates the fit, emphasizing the importance of joint estimation as performed by FACTM (Fig. 2 for $\lambda = 0$ or $\alpha = 10$). Interestingly, for simulations, simple PCA performs relatively well, as the sparsity is not excessively high, while muVI performs suboptimally. In more sparse scenarios, however, PCA's performance deteriorates since it does not account for sparsity, whereas muVI shows even better performance than MOFA (Fig. B.3). Figure B.4 further illustrates the importance of incorporating sparsity-inducing priors in FACTM.

FACTM also outperforms other methods in estimating parameters for structured data (Fig. 3, 4). This advantage is particularly pronounced for the inference of the covariance matrix $\Sigma^{(0)}$, as seen in Figure 4. Figure. B.5 compares $\Sigma^{(0)}$ estimated by FACTM and CTM, showing that, unlike FACTM, CTM fails to capture the true structure of $\Sigma^{(0)}$.

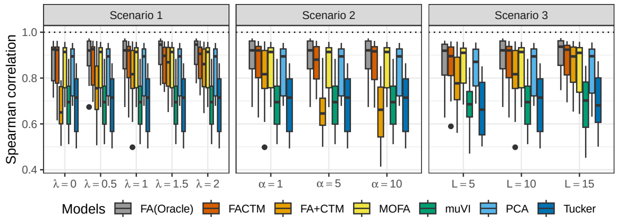

Figure 2: Comparison of true factors and optimally reordered latent factors in factor analysis models.

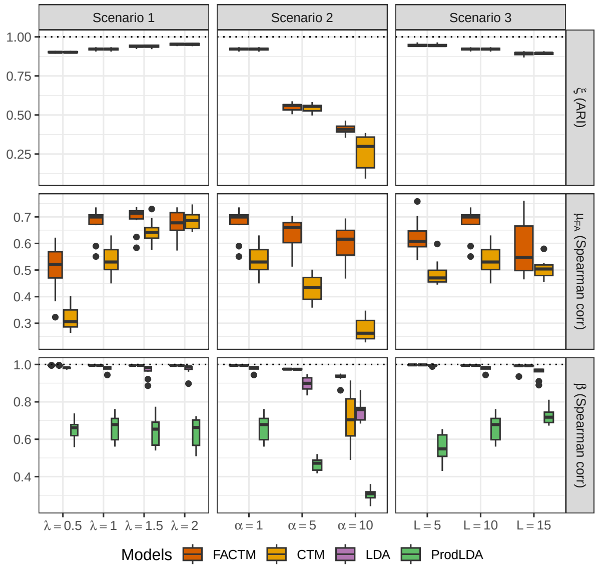

Figure 3: Comparison of true parameters and inferred parameters following the optimal reordering of topics in topic models.

4.2 Predictive power of learned latent representations

We next evaluated the hidden representations inferred by the models on two benchmarks, and one real-world music dataset.

Benchmark datasets We used two multimodal datasets of opinion video clips as benchmarks, with each clip labeled by its sentiment: CMU-MOSI (Zadeh et al., 2016) and CMU-MOSEI (Zadeh et al., 2018). To form simple views we averaged the vision and audio measurements across time stamps, obtaining a single vector per view per sample. For the structured view, we used the transcriptions. Due to the short length of the texts, in FACTM each sentence consisted of a single word for these datasets.

Real-world dataset: Mirex Mirex is a dataset containing songs (Panda et al., 2013). For this dataset, we extracted four simple views: the first containing standard acoustic features extracted using the `pyAudioAnalysis` package (Giannakopoulos, 2015), the second with melody-based features (melody was extracted using the Melodia vamp plug-in (Salamon and Gomez, 2012) with features obtained as in Salamon et al. (2012)), the third with features extracted using the `essentia` package (Bogdanov et al., 2013), and the fourth with common text features like average word length. All the views were quantile normalized. For the structured view, we used the song lyrics, treating each line as a sentence. Each song is labeled with one of five classes: class 1 (described as boisterous, confident, passionate, rousing, rowdy), class 2 (amiable/good natured, cheerful, fun, rollicking, sweet), class 3 (autumnal, bittersweet, brooding, literate, poignant, wistful), class 4 (campy, humorous, silly, witty, wry, whimsical), and class 5 (agressive, fiery, intense, tense/anxious, visceral, volatile). Note that given the descriptions, class 3 is clearly negative, while class 4 is positive. We also retrieved the genre of the songs.

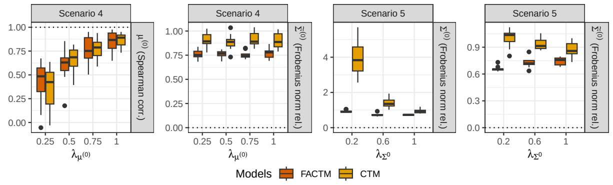

Figure 4: Comparison of true and inferred population-level variables in FACTM and CTM (other topic models do not account for population mean and covariance of topics). The Frobenius distance is computed relative to the Frobenius norm of the true covariance matrix. $\tilde{\Sigma}^{(0)}$ represents the covariance matrix scaled to have ones on the diagonal. Dashed lines indicate optimal performance.

Evaluation To assess the hidden representations learned by the models, we evaluated how informative they were in relation to the assigned labels. Specifically, as sample representations we used either latent factors (Tab. 1), or representations derived from the topics: μ_n for FACTM, $\eta_n - \mu^{(0)}$ for CTM, and log-transformed probabilities for LDA (Tab. 2). For each dataset, we used 10-fold cross-validation to train random forest to predict sample labels using these representations as input features. The classification performance was measured using the AUC for the ROC and Precision-Recall curves (for PR-AUC, see Tab. C.1 and C.2). In the case of CMU-MOSI and CMU-MOSEI, the sentiments were binarized. For the multi-label Mirex dataset, we applied a one-versus-rest approach and calculated the weighted average performance across the classes.

Results Classification on benchmark datasets' representations confirms the simulation results. For the factor-based representations, differences between models are minimal (Tab. 1). For the structured data, FACTM performs best in two cases and consistently outperforms the ablation model - CTM (Tab. 2).

Table 1: Performance of the random forest classifier using inferred latent factors across the datasets. The mean ROC-AUC ± standard deviation values over 10-fold cross-validation are reported.

	Mirex	Mosei	Mosi
FACTM	0.67 ± 0.02	$\mathbf{0.73} \pm 0.01$	$\mathbf{0.68} \pm 0.04$
FA+CTM	$\mathbf{0.68} \pm 0.04$	$\mathbf{0.73} \pm 0.01$	0.67 ± 0.04
MOFA	$\mathbf{0.68} \pm 0.03$	$\mathbf{0.73} \pm 0.01$	0.66 ± 0.04
muVI	$\mathbf{0.68} \pm 0.02$	0.70 ± 0.01	0.67 ± 0.04

Since for the Mirex data, the topic prevalences quantified by FACTM-inferred η_n provided the most predictive representations for class labels, we further explored the insights offered by this representation. Topic 3 shows the second-highest abundance of positive words (Fig. 5A). At the same time, the probability of the lyrics lines (sentences) in songs (samples) from the negative class 3 being assigned to this topic is lower compared to samples from other classes (Fig. 5B). Conversely, sentences in samples from the positive class 4 have a higher probability of being clustered in Topic 4 (in which the number of positive words exceeds the average and negative words are below average, see Fig. 5A and 5C). For graphical representation of word frequencies across all topics, see wordclouds (Fig. B.6).

4.3 Enhancement of interpretability using the rotation method

Further, we showcased the rotation method and the factor interpretability that it brings using the the music dataset Mirex as the testbed.

Table 2: Performance of the random forest classifier using model-specific sample representations across the datasets. The mean ROC-AUC ± standard deviation values over 10-fold cross-validation are reported.

	Mirex	Mosei	Mosi
FACTM	$\mathbf{0.64} \pm 0.06$	0.63 ± 0.01	$\mathbf{0.61} \pm 0.06$
CTM	0.57 ± 0.04	0.55 ± 0.02	0.57 ± 0.05
LDA	0.62 ± 0.05	$\mathbf{0.66} \pm 0.01$	0.57 ± 0.04

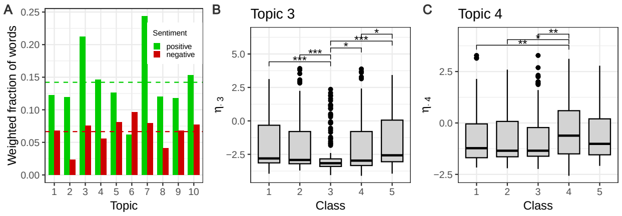

Figure 5: **A**. Topic's average positivity and negativity, measured as the weighted average of positive/negative words in a topic. **B&C**. The values of $\eta_{\cdot,3}$ (**B**) and $\eta_{\cdot,4}$ (**C**) split by class membership of the samples with two-sided Wilcoxon pairwise tests. Stars denote significance of Bonferroni-adjusted p-values: $*** < 0.001$, $** < 0.01$, $* < 0.05$.

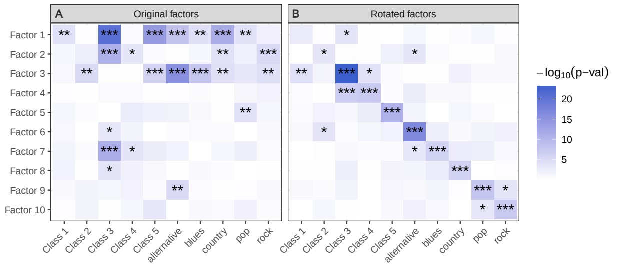

Figure 6: Log-transformed p-values of the two-sided Wilcoxon test for latent factors and binary features (**A**) compared to rotated latent factors and binary features (**B**). Stars denote significance of Bonferroni-adjusted p-values: $*** < 0.001$, $** < 0.01$, $* < 0.05$.

Results To showcase the rotation method, we compared the associations of the originally inferred factors with binary features of the samples in the Mirex dataset to the associations of the rotated factors. For the original factors (Fig. 6A) the binary features have an association mainly with three first factors. After applying our rotation method we obtained many assosiations on the diagonal, improving the interpretability of the consecutive factors (Fig. 6B). Additionally, values of the rotated factors better discriminate between the five classes than the original factors (Fig. B.7). We further show the interpretability of the most discriminative rotated factors by listing their top-weighted features (Fig. B.8).

4.3.1 Ability to infer meaningful topics in real-world, multimodal COVID-19 data

Finally, we demonstrated applicability of FACTM to large scale, real-world multi-modal biological data and its ability to retrieve meaningful clustering in the structured data.

Real-world COVID-19 dataset The dataset consisted of multi-modal data from long COVID-19 patients (Bailey et al., 2024) and included the following modalities: flow cytometry from bronchoalveolar lavage (BAL) fluids, two time-separated sets of CT scans, and scRNA-seq data, which included expression measurements for many single cells per sample. Thus, the dataset consisted of three simple views: one flow cytometry-based view with general cell type fractions per sample, and two CT-based views with the fractions of scans occupied by radiographic abnormalities per sample, as well as one structured view, with each cell treated as a sentence, where gene counts were interpreted as word counts. All simple views underwent quantile normalization, while the structured view was count-normalized, and 1000 highly variable genes were selected for the analysis. We used 20 samples, for which the data for all the views was available, resulting in 169,741 single cells included in the analysis.

Evaluation For this dataset, our main objective was to cluster single cells from the scRNA-seq data and compare FACTM clusters with cell types obtained by scVI method, followed by manual curation, as described in the original article.

Results FACTM identifies biologically relevant clusters in real-world scRNA-seq data (Fig. 7). Indeed, FACTM groups cells into the same cell types as identified in a dedicated clustering and expert annotation step in the original article (Fig. 7A). Even in cases, when one cluster covers more than one cell type, the grouping remains meaningful, e.g. Topic 15 contains cells classified as DC1, DC2, Migratory DC or pDC in the original paper, all of which are subtypes of dendritic cells. Similarly, Topic 16 groups B cells and plasma cells, often referred to as plasma B cells (Fig. B.9). This biological relevance is further supported by the correspondence between gene expression profiles within clusters found by FACTM and the respective cell types, indicating that FACTM learned the correct topic distributions (Fig. 7B).

5 LIMITATIONS AND FUTURE WORK

The proposed model could be further extended in several ways. First, although most hyperparameters in our approach were chosen to result in non-informative priors, or were fixed across all methods for consistent comparison, this may be seen as a limitation of our current study. In future work, we aim to enhance our hyperparameter selection process by incorporating automatic learning techniques, such as the prior predictive matching method proposed in Silva et al. (2023). Secondly, FACTM assumes linear dependencies between the views and latent factors, and the analytical solutions used for optimization in variational inference rely on the conjugacy of parametric distributions, which limits the model's expressiveness. To address this issue, we plan to extend our model to incorporate nonlinear dependencies.

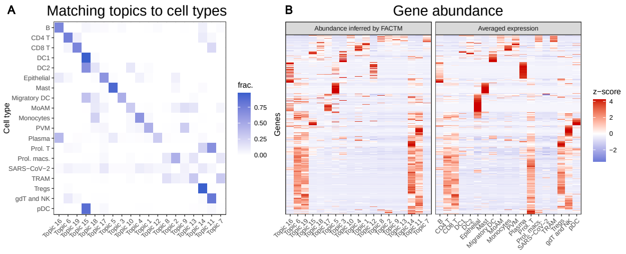

Figure 7: **A**. Contingency table of cell types and clusters inferred by FACTM, scaled such that rows sum to 1. The topics are ordered based on the application of the Hungarian method. **B**. Comparison of z-scored true and inferred average gene expressions for each cell type and each topic. The ordering of the genes in the two panels is the same.

6 CONCLUSIONS

In this work, we addressed the challenge of modeling multi-view and multi-structured data by proposing FACTM, a novel probabilistic Bayesian method that integrates factor analysis with correlated topic model and uses variational inference for optimization. In extensive simulations, FACTM achieved superior accuracy in inferring factors, topics, and their covariance structure. Both factor- and topic-based sample representations learned by FACTM showed state-of-the-art predictive power for label classification on benchmark and real-world music data. On the COVID-19 data, FACTM found meaningful biological clusters. The proposed supervised factor rotation method enhanced factor interpretability for the music dataset. In fact, the rotation method can be used to improve interpretability of any latent factor model. In summary, FACTM adapts the established FA method to handle the complex, structured data of the current world.

Acknowledgements

This work received funding from the Polish National Science Centre SONATA BIS grant No. 2020/38/E/NZ2/00305.

References

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J. C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis - a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124.

Bailey, J. I., Puritz, C. H., Senkow, K. J., Markov, N. S., Diaz, E., Jonasson, E., Yu, Z., Swaminathan, S., Lu, Z., Fenske, S., Grant, R. A., Abdala-Valencia, H., Mylvaganam, R. J., Miller, J., Cumming, R. I., Tighe, R. M., Gowdy, K. M., Kalhan, R., Jain, M., Bharat, A., Kurihara, C., San Jose Estepar, R., San Jose Estepar, R., Washko, G. R., Shilatifard, A., Sznajder, J. I., Ridge, K. M., Budinger, G. S., Braun, R., Misharin, A. V., and Sala, M. A. (2024). Profibrotic monocyte-derived alveolar macrophages are expanded in patients with persistent respiratory symptoms and radiographic abnormalities after COVID-19. *Nature immunology*, 25(11):2097–2109.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P., and Goodman, N. D. (2019). Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20:973 – 978.

Blei, D., Kucukelbir, A., and McAuliffe, J. (2016). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of Science. *The Annals of Applied Statistics*, 1(1):17 – 35.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Bogdanov, D., Wack, N., Gómez, E., Gulati, S., Herrera, P., Mayor, O., Roma, G., Salamon, J., Zapata, J., and Serra, X. (2013). Essentia: an audio analysis library for music information retrieval. In *14th International Society for Music Information Retrieval Conference*.

Bredikhin, D., Kats, I., and Stegle, O. (2022). MUON: multimodal omics analysis framework. *Genome Biology*, 23(42).

Chen, Z., Soifer, I., Hilton, H. G., Keren, L., and Jojic, V. (2020). Modeling Multiplexed Images with Spatial-LDA Reveals Novel Tissue Microenvironments. *Journal of Computational Biology*, 27:1204 – 1218.

Fallah, K. and Rozell, C. J. (2022). Variational sparse coding with learned thresholding. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 6034–6058. PMLR.

Giannakopoulos, T. (2015). pyAudioAnalysis: An Open-Source Python Library for Audio Signal Analysis. *PLoS ONE*, 10(12).

Kabsch, W. (1976). A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32:922–923.

Kaiser, H. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 23(3):187–200.

Klami, A., Virtanen, S., Leppäaho, E., and Kaski, S. (2014). Group Factor Analysis. *IEEE transactions on neural networks and learning systems*, 26.

Kossaifi, J., Panagakis, Y., Anandkumar, A., and Pantic, M. (2019). TensorLy: Tensor Learning in Python. *Journal of Machine Learning Research*, 20(26):1–6.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52.

Lafferty, J. and Blei, D. (2005). Correlated Topic Models. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press.

Masada, T. and Takasu, A. (2013). A revised inference for correlated topic model. In Guo, C., Hou, Z.-G., and Zeng, Z., editors, *Advances in Neural Networks – ISNN 2013*, pages 445–454, Berlin, Heidelberg. Springer Berlin Heidelberg.

Miller, B., Huang, F., Atta, L., Sahoo, A., and Fan, J. (2022). Reference-free cell type deconvolution of multi-cellular pixel-resolution spatially resolved transcriptomics data. *Nature Communications*, 13.

Panda, R., Malheiro, R., Rocha, B., Oliveira, A., and Paiva, R. P. (2013). Multi-modal music emotion recognition: A new dataset, methodology and comparative analysis. In *10th International Symposium on Computer Music Multidisciplinary Research*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Qoku, A. and Buettner, F. (2023). Encoding Domain Knowledge in Multi-view Latent Variable Models: A Bayesian Approach with Structured Sparsity. In

Ruiz, F., Dy, J., and van de Meent, J.-W., editors, *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 11545–11562. PMLR.

Salamon, J. and Gomez, E. (2012). Melody extraction from polyphonic music signals using pitch contour characteristics. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(6):1759–1770.

Salamon, J., Rocha, B., and Gómez, E. (2012). Musical genre classification using melody features extracted from polyphonic music signals. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Silva, E. d. S. d., Kuśmierczyk, T., Hartmann, M., and Klami, A. (2023). Prior specification for Bayesian matrix factorization via prior predictive matching. *Journal of Machine Learning Research*, 24(67):1–51.

Srivastava, A. and Sutton, C. (2017). Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Thurstone, L. L. (1931). Multiple factor analysis. *Psychological Review*, 38(5):406–427.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622.

Tonolini, F., Jensen, B. S., and Murray-Smith, R. (2020). Variational sparse coding. In Adams, R. P. and Gogate, V., editors, *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pages 690–700. PMLR.

Velten, B., Braunger, J. M., Argelaguet, R., Arnol, D., Wirbel, J., Bredikhin, D., Zeller, G., and Stegle, O. (2022). Identifying temporal and spatial patterns of variation from multimodal data using MEFISTO. *Nature Methods*, 19(2):179–186.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Zadeh, A., Liang, P. P., Vanbriesen, J., Poria, S., Tong, E., Cambria, E., Minghai, C., and Morency, L.-P. (2018). Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In Gurevych, I. and Miyao, Y., editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Zadeh, A., Zellers, R., Pincus, E., and Morency, L.-P. (2016). MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*.

Zhao, S., Gao, C., Mukherjee, S., and Engelhardt, B. E. (2016). Bayesian group factor analysis with structured sparsity. *Journal of Machine Learning Research*, 17(196):1–47.

Checklist

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

   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes/No/Not Applicable]
   See Section 3.2 for model description and Section 3.3 for optimization details.

   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes/No/Not Applicable]
   We compare computational complexity of our model to existing methods in Section 3.3.

   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes/No/Not Applicable]
   We will provide a link to the GitHub repository after the reviewing process is complete and the anonymity requirements are lifted.

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

   (a) Statements of the full set of assumptions of all theoretical results. [Yes/No/Not Applicable]

   (b) Complete proofs of all theoretical results. [Yes/No/Not Applicable]

   (c) Clear explanations of any assumptions. [Yes/No/Not Applicable]

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

   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes/No/Not Applicable]
   The code is available via the GitHub repository, which will be made accessible as soon as the anonymity requirements are lifted.

   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes/No/Not Applicable]

   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [[✓ Yes]/No/Not Applicable]

   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes/[✓ No]/Not Applicable]

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

   (a) Citations of the creator If your work uses existing assets. [[✓ Yes]/No/Not Applicable]

   (b) The license information of the assets, if applicable. [[✓ Yes]/No/Not Applicable]
   See Section 4, paragraph on 'Compared Methods'. We will include the licence information in the GitHub repository.

   (c) New assets either in the supplemental material or as a URL, if applicable. [[✓ Yes]/No/Not Applicable]
   We will upload the code to the GitHub repository, which will be made accessible as soon as the anonymity requirements are lifted.

   (d) Information about consent from data providers/curators. [Yes/No/[✓ Not Applicable]]
   We use publicly available datasets.

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Yes/No/[✓ Not Applicable]]

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

   (a) The full text of instructions given to participants and screenshots. [Yes/No/[✓ Not Applicable]]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Yes/No/[✓ Not Applicable]]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Yes/No/[✓ Not Applicable]]

A DERIVATIONS

FACTM combines factor analysis (FA) with correlated topic model (CTM), and while many of the variational update equations follow the ones for the original models (specifically, Argelaguet et al. (2018) for the FA and Lafferty and Blei (2005) for the CTM part), our extension required modifications and additional updates.

In Section A.1, we present the update formulas of variational parameters in coordinate ascent algorithm for the variable μ_n, which links the FA and CTM parts of the model, as well as for the CTM adapted for sentence-based structured data. While the latter represents a relatively small extension to the standard CTM (Lafferty and Blei, 2005), we provide the equations for completeness. The key difference lies in the fact that sentences introduce a constraint, where each data point (sentence), composed of a set of observed objects (words), must originate from a single topic. The original CTM becomes a special case of our model, where there is one word per sentence. On the one hand, this special case is most flexible, allowing for a different distribution of topics for each observed object (word). On the other hand, our extension accounts for a reasonable assumption that each entire sentence groups words of the same topic. Moreover, in some applications assuming multi-object data point is crucial for capturing the nature of the data (see e.g. COVID-19 example in the main text, where sentences correspond to single cells, words to measured genes and the topics correspond to cell types).

Section A.2 addresses the adjustments in the update formulas for the FA part in FACTM.

A.1 Variational inference for the structured model based on CTM with sentences

Overview of notation and assumptions Recall from the main text, that the ELBO is defined as

$$ELBO(q) := \mathbb{E}_q \log p(X,Y) - \mathbb{E}_q \log q(X) = \mathbb{E}_q \log p(X,Y) + H(q). \tag{4}$$

In this equation, q represents a variational distribution, while X and Y hidden and observed variables, respectively. The term $-\mathbb{E}_q \log q(X)$ in (4) corresponds to the entropy denoted by H of the distribution q. The optimal variational distribution $q_i := q(X_i)$ is determined by the expression:

$$\log q(X_i) \propto \mathbb{E}_{-q_i} \log p(X,Y), \tag{5}$$

where the expectation is taken over all variational distributions except q_i corresponding to the variable X_i. Alternatively, assuming a parametric form for the variational distribution, the ELBO can also be optimized directly.

We simplify the joint distribution of the FACTM model showed in Figure 1 from the main text by focusing only on the structured component. For this consideration, we assume that the nodes related to the factor analysis part, $z_{n,k}$ and $\overline{w}_{l,k}$ for $n \in \{1,2,\ldots,N\}$, $k \in \{1,2,\ldots,K\}$, $l \in \{1,2,\ldots,L\}$, are fixed (note that this simplification does not affect the correctness of the update equations in the full model). Thus, based on Equation (2) from the main text, the joint probability density we consider is given by

$$\begin{aligned} p(\overline{W},\mu,\eta,\xi,\beta,\overline{Y}|\mu^{(0)},\Sigma^{(0)},t,\alpha_0^\beta) = & \prod_{n=1}^{N}\prod_{l=1}^{L}\mathcal{N}(\mu_{n,l}|\sum_{k=1}^{K} z_{n,k}\overline{w}_{l,k}, 1/t)\prod_{n=1}^{N}\mathcal{N}_L(\eta_n|\mu_n+\mu^{(0)},\Sigma^{(0)}) \\ & \prod_{n=1}^{N}\prod_{i=1}^{I}\mathrm{Mult}(\xi_{n,i}|1,\mathrm{softmax}(\eta_n))\prod_{l=1}^{L}\mathrm{Dir}(\beta_l|\alpha_0^\beta) \\ & \prod_{n=1}^{N}\prod_{i=1}^{I}\prod_{j=1}^{J_i}\mathrm{Mult}(\overline{y}_{n,i,j}|1,\beta_{\xi_{n,i}}) \end{aligned}$$

where $\mathrm{softmax}(\eta_n) = (\exp(\eta_{n,1})/C, \exp(\eta_{n,2})/C, \ldots, \exp(\eta_{n,L})/C)$ and $C = \sum_{l=1}^{L}\exp(\eta_{n,l})$. In this scenario, the variational distribution, assuming a mean-field approximation along with a parametric form for $q(\eta_{n,l})$ and $q(\xi_{n,i})$ as in the main text (see Equation (3) there), takes the following form, incorporating the variational parameters:

$$q(\mu,\eta,\xi,\beta) = \prod_{n=1}^{N} q_L(\mu_n|\mu_n^{(\mu)},\Sigma_n^{(\mu)})\prod_{n=1}^{N}\prod_{l=1}^{L}\mathcal{N}(\eta_{n,l}|\mu_{n,l}^{(\eta)},(\sigma_{n,l}^{(\eta)})^2)\prod_{n=1}^{N}\prod_{i=1}^{I_n}\mathrm{Mult}(\xi_{n,i}|1,\phi_{n,i})\prod_{l=1}^{L} q(\beta_l|\alpha_l^{(\beta)}).$$

Thus a formula for the ELBO for this specific model is given by

$$ELBO(q) = \sum_{n=1}^{N} \Big(\sum_{l=1}^{L} \mathbb{E}_q \log \mathcal{N}(\mu_{n,l} | \sum_{k=1} z_{n,k} \overline{w}_{l,k}, 1/t) \tag{6a}$$

$$+ \mathbb{E}_q \log \mathcal{N}_L(\eta_n | \mu_n + \mu^{(0)}, \Sigma^{(0)}) \tag{6b}$$

$$+ \sum_{i=1}^{I_n} \mathbb{E}_q \log \mathrm{Mult}(\xi_{n,i} | 1, \mathrm{softmax}(\eta_n)) \tag{6c}$$

$$+ \sum_{i=1}^{I_n} \sum_{j=1}^{J_i} \mathbb{E}_q \log \mathrm{Mult}(\overline{y}_{n,i,j} | 1, \beta_{\xi_{n,i}}) \Big) \tag{6d}$$

$$+ \sum_{l=1}^{L} \mathbb{E}_q \log \mathrm{Dir}(\beta_l | \alpha) \tag{6e}$$

$$+ H(q_L(\mu_n)) + H(q(\eta_{n,l})) + H(q(\xi_{n,i})) + H(q(\beta_l)). \tag{6f}$$

The entropy of q in (6f) for the variables, for which the parametric distribution is assumed, equals

$$H(q(\eta_{n,l})) = \log((\sigma_{n,l}^{(\eta)})^2)/2 + \log(\sqrt{2\pi e}) \tag{7}$$

and

$$H(q(\xi_{n,i})) = \sum_{l=1}^{L} \phi_{n,i,l} \log(\phi_{n,i,l}). \tag{8}$$

Now we obtain the update equations for all the variational parameters, all of which are listed below, including the parameters for the link variable μ_n. Additionally, we provide also updates for $\mu^{(0)}$ and $\Sigma^{(0)}$.

Variational parameters in the structered part of the model

- μ_n for $n = 1, 2, \ldots, N$
  Variational parameters: $\mu_n^{(\mu)}$, $\Sigma_n^{(\mu)}$
- η_n for $n = 1, 2, \ldots, N$
  Variational parameters: $\mu_n^{(\eta)}$, $(\sigma_n^{(\eta)})^2$, and additional parameter ζ_n introduced in Section A.1.1
- $\xi_{n,i}$ for $n = 1, 2, \ldots, N$, $i = 1, 2, \ldots, I_n$
  Variational parameters: $\phi_{n,i}$
- β_l for $l = 1, 2, \ldots, L$
  Variational parameters: $\alpha_l^{(\beta)}$

A.1.1 The update equations

For clarity in the following sections, we introduce the notation $\tilde{\xi}_{n,i}$. Recall that $\xi_{n,i}$ is clustering variable that takes a value from the set of topics $\{1, 2, \ldots, L\}$. The variable $\tilde{\xi}_{n,i}$ is a one-hot encoded vector, where all elements are zero except for the lth position, which is set to 1 if $\xi_{n,i} = l$.

- μ_n for $n = 1, 2, \ldots, N$
  We provide a detailed derivation of the updates for $\mu_n^{(\mu)}$ and $\Sigma_n^{(\mu)}$ from Section 3.3 of the main text. Using (5), rather than optimizing the ELBO directly, we can compute the following expected value:

$$\log(q(\mu_n)) \propto \mathbb{E}_q \log \big(\mathcal{N}_L(\mu_n | \sum_{k=1}^{K} z_{n,k} \overline{w}_{\cdot,k}, \mathrm{diag}(1/t)) \cdot \mathcal{N}_L(\eta_n | \mu_n + \mu^{(0)}, \Sigma^{(0)}) \big),$$

where $\mathbb{E}_{-\mu_n}$ denotes the expectation taken over all variables following the distribution q, except for μ_n. If we interpret the second normal density as $\mathcal{N}_L(\eta_n - \mu^{(0)} | \mu_n, \Sigma^{(0)})$, and treat the first normal distribution

$\mathcal{N}_L(\mu_n|\sum_{k=1}^K z_{n,k}\overline{w}_{\cdot,k}, \mathrm{diag}(1/t))$ as the prior for μ_n and the second $\mathcal{N}_L(\eta_n - \mu^{(0)}|\mu_n, \Sigma^{(0)})$ as the likelihood, then the posterior distribution is also normal $\mathcal{N}_L(\mu_n|\tilde{\mu}, \tilde{\Sigma})$, with parameters given by

$$\tilde{\Sigma} = \left(\mathrm{diag}(t) + \left(\Sigma^{(0)}\right)^{-1}\right)^{-1}$$

and

$$\tilde{\mu} = \tilde{\Sigma}\left(\mathrm{diag}(t)\sum_{k=1}^{K} z_{n,k}\overline{w}_{\cdot,k} + \left(\Sigma^{(0)}\right)^{-1}(\eta_n - \mu^{(0)})\right).$$

Thus, after applying expected value, we get that the updates of the parameters of $q_L(\mu_n)$ are

$$\hat{\Sigma}_n^{(\mu)} = \left(\mathrm{diag}(t) + \left(\Sigma^{(0)}\right)^{-1}\right)^{-1},$$

$$\hat{\mu}_n^{(\mu)} = \hat{\Sigma}_n^{(\mu)}\Big(\mathrm{diag}(t)\sum_{k=1}^{K} z_{n,k}\overline{w}_{\cdot,k} + \left(\Sigma^{(0)}\right)^{-1}(\mathbb{E}_q\eta_n - \mu^{(0)})\Big),$$

where $\mathbb{E}_q\eta_n = \mu_n^{(\eta)}$. We recall that for now, focusing solely on the structured part, we have assumed that both $z_{n,k}$ and $\overline{w}_{\cdot,k}$ are deterministic. However, in the complete FACTM, these are treated as random variables, and their expectations must also be computed.

- η_n for $n = 1, 2, \ldots, N$
  As our objective is to maximize the *ELBO* in equations (6a) - (6f) with respect to variational parameters of η_n, the relevant terms involving η_n are (6b), (6c), and the corresponding entropy term in (6f). Since direct optimization is infeasible, we derive a lower bound as in Lafferty and Blei (2005) (up to constant terms with respect to parameters $\mu_n^{(\eta)}$ and $(\sigma_n^{(\eta)})^2$) by introducing a new auxiliary parameter $\zeta_n > 0$. Since there is no analytic solution for maximizing the lower bound with respect to $\mu_n^{(\eta)}$ and $(\sigma_n^{(\eta)})^2$, we compute the gradient and employ the L-BFGS-B algorithm for optimization.

  First, we focus on the term (6c). This gives

$$\mathbb{E}_q \log \mathrm{Mult}(\xi_{n,i}|1, \mathrm{softmax}(\eta_n)) \propto \mathbb{E}_q \log\left(\frac{\exp(\eta_n'\tilde{\xi}_{n,i})}{\sum_{l=1}^{L}\exp(\eta_{n,l})}\right) = \mathbb{E}_q\eta_n'\tilde{\xi}_{n,i} - \mathbb{E}_q\log\left(\sum_{l=1}^{L}\exp(\eta_{n,l})\right), \tag{9}$$

  where $\propto$ indicates equality up to terms constant with respect to $\mu_n^{(\eta)}$ and $(\sigma_n^{(\eta)})^2$. Next, we apply a Taylor expansion of the logarithm around the auxiliary parameter ζ_n

$$\log\left(\sum_{l=1}^{L}\exp(\eta_{n,l})\right) = \log(\zeta_n) + \sum_{m=1}^{\infty}\frac{(-1)^{m-1}}{m\zeta_n^m}\left(\sum_{l=1}^{L}\exp(\eta_{n,l}) - \zeta_n\right)^m$$

  to upper bound the second term of Equation (9)

$$\mathbb{E}_q\log\left(\sum_{l=1}^{L}\exp(\eta_{n,l})\right) \leq \log(\zeta_n) + \zeta_n^{-1}\sum_{l=1}^{L}\mathbb{E}_q\exp(\eta_{n,l}) - 1.$$

  Summing over $i \in \{1, 2, \ldots, I_n\}$, we lower bound term (6c) in the following way

$$\sum_{i=1}^{I}\mathbb{E}_q\log\mathrm{Mult}(\tilde{\xi}_{n,i}|1, \mathrm{softmax}(\eta_n)) \geq \sum_{i=1}^{I}\mathbb{E}_q\eta_n'\tilde{\xi}_{n,i} - I\big(\log(\zeta_n) + \zeta_n^{-1}\sum_{l=1}^{L}\mathbb{E}_q\exp(\eta_{n,l}) - 1\big). \tag{10}$$

  The expected values in (10) equal

$$\mathbb{E}_q\eta_n'\tilde{\xi}_{n,i} = \sum_{l=1}^{L}\mu_{n,l}^{(\eta)}\phi_{n,i,l} \tag{11}$$

and

$$\mathbb{E}_q \exp(\eta_{n,l}) = \exp\left(\mu^{(\eta)}_{n,l} + (\sigma^{(\eta)}_{n,l})^2/2\right), \tag{12}$$

where we use the fact, that the second expectation is the moment generating function of a normal distribution evaluated at 1. Next, we expand the term (6b)

$$\mathbb{E}_q \log \mathcal{N}_l(\eta_n|\mu_n+\mu^{(0)},\Sigma^{(0)}) \propto -\mathbb{E}_q(\eta_n-\mu_n-\mu^{(0)})'\left(\Sigma^{(0)}\right)^{-1}(\eta_n-\mu_n-\mu^{(0)})/2, \tag{13}$$

and compute the expected value of a quadratic form

$$\begin{aligned}\mathbb{E}_q(\eta_n-\mu_n-\mu^{(0)})'\left(\Sigma^{(0)}\right)^{-1}(\eta_n-\mu_n-\mu^{(0)}) = \operatorname{tr}\left(\left(\operatorname{diag}((\sigma^{(\eta)}_n)^2)+\Sigma^{(\mu)}_n\right)\left(\Sigma^{(0)}\right)^{-1}\right)& \\ + (\mu^{(\eta)}_n-\mu^{(\mu)}_n-\mu^{(0)})'\left(\Sigma^{(0)}\right)^{-1}(\mu^{(\eta)}_n-\mu^{(\mu)}_n-\mu^{(0)}),& \end{aligned} \tag{14}$$

where $\operatorname{diag}((\sigma^{(\eta)}_n)^2)$ denotes an $L\times L$ diagonal matrix with $(\sigma^{(\eta)}_{n,l})^2$ as its diagonal elements. Thus, by combining the results from (6b), (6c), and the entropy from (7), we get the lower bound that needs to be maximized with respect to $(\mu^{(\eta)}_n,(\sigma^{(\eta)}_n)^2)$, and ζ_n:

$$\begin{aligned}f(\mu^{(\eta)}_n,(\sigma^{(\eta)}_n)^2,\zeta_n) = -\operatorname{tr}\left(\operatorname{diag}\left((\sigma^{(\eta)}_n)^2\right)\left(\Sigma^{(0)}\right)^{-1}\right)/2 - (\mu^{(\eta)}_n-\mu^{(\mu)}_n-\mu^{(0)})'\left(\Sigma^{(0)}\right)^{-1}(\mu^{(\eta)}_n-\mu^{(\mu)}_n-\mu^{(0)})/2 \\ + \sum_{l=1}^{L}\mu^{(\eta)}_{n,l}\sum_{i=1}^{I}\phi_{n,i,l} - I\left(\log(\zeta_n)+\zeta_n^{-1}\sum_{l=1}^{L}\exp\left(\mu^{(\eta)}_{n,l}+(\sigma^{(\eta)}_{n,l})^2/2\right)\right) + \sum_{l=1}^{L}\log((\sigma^{(\eta)}_{n,l})^2)/2.\end{aligned}$$

Gradients of f equal

$$\nabla_{\mu^{(\eta)}_n} f(\mu^{(\eta)}_n,(\sigma^{(\eta)}_n)^2,\zeta_n) = -\left(\Sigma^{(0)}\right)^{-1}(\mu^{(\eta)}_n-\mu^{(\mu)}_n-\mu^{(0)}) + \sum_{i=1}^{I}\phi_{n,i} - I\zeta_n^{-1}\exp\left(\mu^{(\eta)}_n+(\sigma^{(\eta)}_n)^2/2\right),$$

$$\nabla_{(\sigma^{(\eta)}_n)^2} f(\mu^{(\eta)}_n,(\sigma^{(\eta)}_n)^2,\zeta_n) = -\operatorname{diag}\left(\left(\Sigma^{(0)}\right)^{-1}\right)/2 - I\zeta_n^{-1}\exp\left(\mu^{(\eta)}_n+(\sigma^{(\eta)}_n)^2/2\right)/2 + 1/(2(\sigma^{(\eta)}_n)^2).$$

All terms in the equations above are vectors of length L, and the transformations are applied elementwise (e.g. $\exp(\mu^{(\eta)}_n) = (\exp(\mu^{(\eta)}_{n,1}),\exp(\mu^{(\eta)}_{n,2}),\dots,\exp(\mu^{(\eta)}_{n,L}))$. The optimal value of ζ_n obtained analytically from f equals

$$\hat{\zeta}_n = \sum_{l=1}^{L}\exp(\mu^{(\eta)}_{n,l}+(\sigma^{(\eta)}_{n,l})^2/2).$$

- $\xi_{n,i}$ for $n=1,2,\dots,N$, $i=1,2,\dots,I_n$
  As the terms of ELBO involving $\xi_{n,i}$ are (6c), (6d) and the corresponding entropy term in (6f) (computed in (8)), for fixed n,i,l the function we aim to maximize with respect to $\phi_{n,i,l}$ up to constants is

$$\begin{aligned}f(\phi_{n,i,l}) &= \mathbb{E}_q\eta_{n,l}\tilde{\xi}_{n,i,l} + \sum_{j=1}^{J_i}\mathbb{E}_q\left(\tilde{\xi}_{n,i,l}\log\beta_{l,y_{n,i,j}}\right) - \phi_{n,i,l}\log\phi_{n,i,l} \\ &= \mu^{(\eta)}_{n,l}\phi_{n,i,l} + \phi_{n,i,l}\sum_{j=1}^{J_i}\mathbb{E}_q\log\beta_{l,y_{n,i,j}} - \phi_{n,i,l}\log\phi_{n,i,l}.\end{aligned}$$

This results in

$$\hat{\phi}_{n,i,l} \propto \exp\left(\mu^{(\eta)}_{n,l} + \sum_{j=1}^{J_i}\mathbb{E}_q\log\beta_{l,\tilde{y}_{n,i,j}}\right),$$

where $\hat{\phi}_{n,i,l}$ is known up to multiplicative constants. That issue is solved by noting that $\sum_{l=1}^{L}\phi_{n,i,l}=1$, and after rescaling, we obtain the final value for $\hat{\phi}_{n,i,l}$.

Now we compute the term $\sum_{j=1}^{J_i} \mathbb{E}_q \log \beta_{l,\tilde{y}_{n,i,j}}$. Note that

$$\sum_{j=1}^{J_i} \mathbb{E}_q \log \beta_{l,\nu_{n,i,j}} = \sum_{g=1}^{G} \underbrace{\sum_{j=1}^{J_i} \mathbb{I}(\overline{y}_{n,i,j} = g)}_{\substack{\text{number of words } g \\ \text{in a sentence } i}} \mathbb{E}_q \log \beta_{l,g} = \sum_{g=1}^{G} \overline{y}_{n,i,g} \mathbb{E}_q \log \beta_{l,g}.$$

As a variational distribution of β_l is Dirichlet$(\alpha_l^{(\beta)})$, we have that β_{l,g_0} is Beta$(\alpha_{l,g_0}^{(\beta)}, \sum_{g \neq g_0} \alpha_{l,g}^{(\beta)})$ and hence we obtain

$$\mathbb{E}_q \log \beta_{l,g_0} = \psi(\alpha_{l,g_0}^{(\beta)}) - \psi\left(\sum_{g=1}^{G} \alpha_{l,g}^{(\beta)}\right), \tag{15}$$

where ψ is digamma function.

- β_l for $l = 1, 2, \ldots, L$

  For β_l, we assume conjugate distributions, where the prior is a Dirichlet distribution and the likelihood follows a Multinomial distribution, thus we simply have that the optimal $q(\beta_l)$ is also Dirichlet with parameters

$$\alpha_{l,g}^{(\beta)} = \alpha_0^{\beta} + \sum_{n=1}^{N} \sum_{i=1}^{I_n} \phi_{n,i,l} \overline{y}_{n,i,g}.$$

- $\mu^{(0)}$ and $\Sigma^{(0)}$

  To get updates of $\mu^{(0)}$ and $\Sigma^{(0)}$, we need to maximize a function (see (13) and (14))

$$\begin{aligned}
&\sum_{n=1}^{N} \Big(- \log(\det(\Sigma^{(0)}))/2 - \operatorname{tr}\left(\left(\operatorname{diag}((\sigma_n^{(\eta)})^2) + \Sigma_n^{(\mu)}\right) \left(\Sigma^{(0)}\right)^{-1}\right)/2 \\
&\qquad\qquad - (\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})' \left(\Sigma^{(0)}\right)^{-1} (\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})/2\Big) \\
&\qquad\qquad = \sum_{n=1}^{N} \Big(- \log(\det(\Sigma^{(0)}))/2 - \operatorname{tr}\big(\big(\operatorname{diag}((\sigma_n^{(\eta)})^2) + \Sigma_n^{(\mu)}) \\
&\qquad\qquad\qquad\qquad + (\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})(\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})'\big) \left(\Sigma^{(0)}\right)^{-1})/2\Big).
\end{aligned}$$

This task is equivalent to finding ML estimators of the parameters of multivariate normal distribution, thus

$$\mu^{(0)} = \frac{1}{n} \sum_{n=1}^{N} \left(\mu_n^{(\eta)} - \mu_n^{(\mu)}\right),$$

$$\Sigma^{(0)} = \frac{1}{n} \sum_{n=1}^{N} \left(\Sigma^{(\mu)} + \operatorname{diag}((\sigma_n^{(\eta)})^2) + (\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})(\mu_n^{(\eta)} - \mu_n^{(\mu)} - \mu^{(0)})'\right).$$

Remark It is important to note that the update equations do not require the counts $\overline{y}_{i,j,g}$ to be integers. Instead, they can also be interpreted as weights, as long as they remain positive.

A.2 Variational inference for the FA part

For the FA part, we provide only the final update equations for $\overline{w}_{l,k}$ and $z_{n,k}$, as their derivations follow standard procedures, such as those in Argelaguet et al. (2018), and can be easily obtained from Equation (1) from the main text. The key modification is that unobserved $y_{n,l}$ is replaced by the expectation of the link variable $\mu_{n,l}$, namely $\mathbb{E}_q \mu_{n,l} = \mu_{n,l}^{(\mu)}$. Additionally, the data precision τ is substituted by parameter t for structured views, and we do not use spike-and-slab prior for loadings $\overline{w}$ in these structured view.

Selected variational parameters in FA part of the model

- $\overline{w}_{l,k}$ for $l = 1, 2, \ldots, L$, $k = 1, 2, \ldots, K$
  Variational parameters: $\mu^{(\overline{w})}_{l,k}$, $(\sigma^{(\overline{w})}_{l,k})^2$
- $z_{n,k}$ for $n = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$
  Variational parameters: $\mu^{(z)}_{n,k}$, $(\sigma^{(z)}_{n,k})^2$

A.3 The update equations

- $\overline{w}_{l,k}$ for $l = 1, 2, \ldots, L$, $k = 1, 2, \ldots, K$
  The variational distribution of $\overline{w}_{l,k}$ is a normal distribtuion and the updates for $\mu^{(\overline{w})}_{l,k}$ and $(\sigma^{(\overline{w})}_{l,k})^2$ are

$$\hat{\mu}^{(\overline{w})}_{l,k} = t(\hat{\sigma}^{(\overline{w})}_{l,k})^2 \left(\sum_{n=1}^{N} \mathbb{E}_q z_{n,k} \left(\mathbb{E}_q \mu_{n,l} - \sum_{k' \neq k} \mathbb{E}_q \overline{w}_{l,k'} \mathbb{E}_q z_{n,k'} \right) \right),$$

$$(\hat{\sigma}^{(\overline{w})}_{l,k})^2 = \left(t \sum_{n=1}^{N} \mathbb{E}_q z^2_{n,k} + \mathbb{E}_q \overline{\alpha}_k \right)^{-1},$$

  where $\mathbb{E}_q \mu_{n,l} = \mu^{(\mu)}_{n,l}$, $\mathbb{E}_q \overline{w}_{l,k} = \mu^{(\overline{w})}_{l,k}$, and the remaining expected values are analogous to those in Argelaguet et al. (2018).

- $z_{n,k}$ for $n = 1, 2, \ldots, N$, $k = 1, 2, \ldots, K$
  The variational distribution of $z_{n,k}$ a normal distribution and the updates for $\mu^{(z)}_{n,k}$ and $(\sigma^{(z)}_{n,k})^2$ are

$$\begin{aligned}\hat{\mu}^{(z)}_{n,k} = (\hat{\sigma}^{(z)}_{n,k})^2 &\left(\sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q \tau^m_d \mathbb{E}_q w^m_{d,k} \left(y^m_{n,d} - \sum_{k' \neq k} \mathbb{E}_q z_{n,k'} \mathbb{E}_q w^m_{d,k'} \right) \right. \\ &\left. \sum_{l=1}^{L} t \mathbb{E}_q \overline{w}_{l,k} \left(\mathbb{E}_q \mu_{n,l} - \sum_{k' \neq k} \mathbb{E}_q z_{n,k'} \mathbb{E}_q \overline{w}_{l,k'} \right) \right),\end{aligned}$$

$$(\hat{\sigma}^{(z)}_{n,k})^2 = \left(1 + \sum_{m=1}^{M} \sum_{d=1}^{D_m} \mathbb{E}_q (w^m_{d,k})^2 + \sum_{l=1}^{L} \mathbb{E}_q (\overline{w}_{l,k})^2 \right)^{-1},$$

  where $\mathbb{E}_q \mu_{n,l} = \mu^{(\mu)}_{n,l}$, $\mathbb{E}_q \overline{w}_{l,k} = \mu^{(\overline{w})}_{l,k}$, $\mathbb{E}_q (w^m_{d,k})^2 = (\mu^{(\overline{w})}_{l,k})^2 + (\sigma^{(\overline{w})}_{l,k})^2$, and the remaining expected values are analogous to those in Argelaguet et al. (2018).

B ADDITIONAL FIGURES

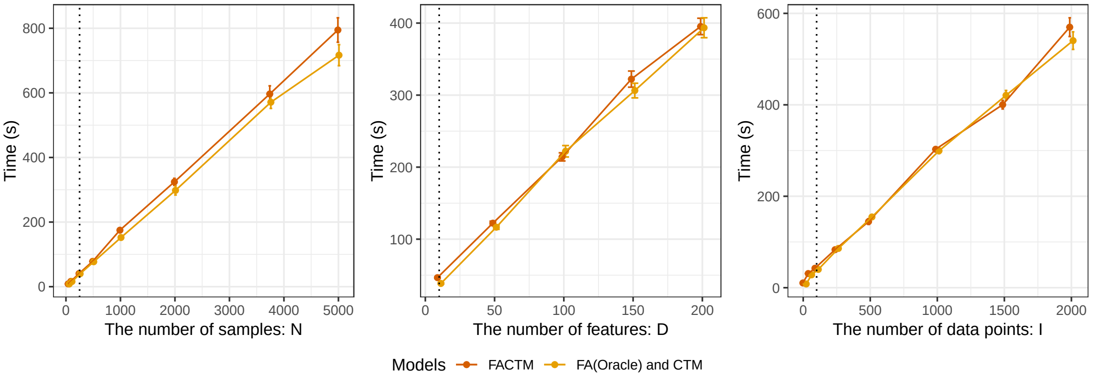

Figure B.1: Runtime analysis using the baseline simulation setup (marked with vertical lines) showing scalability dependence on key data parameters: N, D, and I. The execution time of FACTM is compared to the combined runtime of its individual components, FA (FA Oracle) and CTM, over 10 iterations of each algorithm across 5 model fits on 10 datasets. For each dataset, the results of the 5 models were averaged. Data points represent these averages, with error bars showing the standard deviation across datasets.

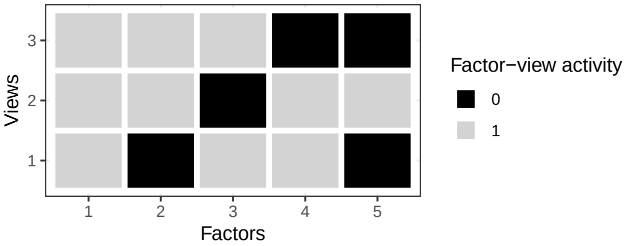

Figure B.2: Illustration of factor-wise sparsity used in simulations, where 0 indicates an inactive factor for a given view, and 1 indicates an active factor.

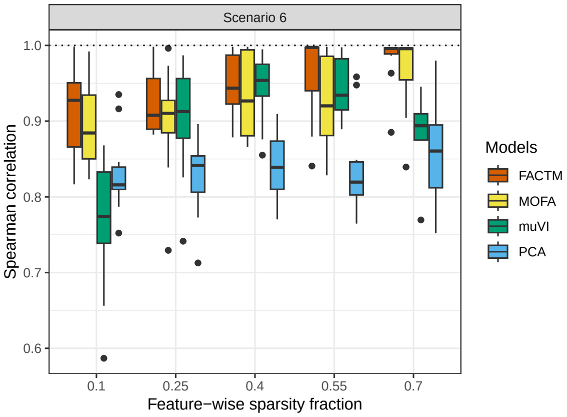

Figure B.3: Comparison of true factors and optimally reordered latent factors for muVI, MOFA and PCA in an additional simulation scenario (Scenario 6), which explores varying levels of feature-wise sparsity.

Scenario 6: This scenario is based on our standard setup, with a few modifications. We increased the number of features in both simple views to $D = 500$ (to support the inference of sparsity-related parameters for both models). The basic feature-wise sparsity fraction, initially set at 0.1, was increased to four values: $\{0.25, 0.4, 0.55, 0.7\}$.

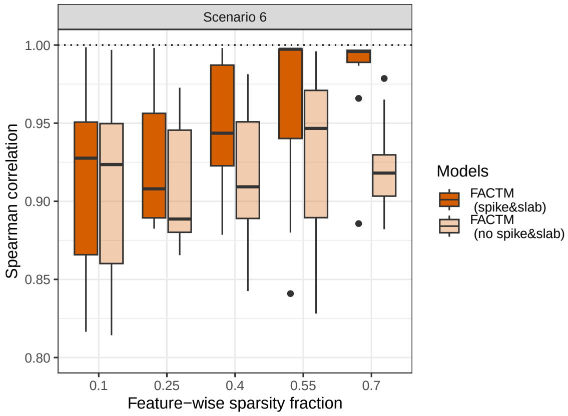

Figure B.4: Comparison of true factors and optimally reordered latent factors for FACTM with and without spike and slab prior in an additional simulation scenario (Scenario 6), which explores varying levels of feature-wise sparsity.
For the description of Scenario 6 see Figure B.3.

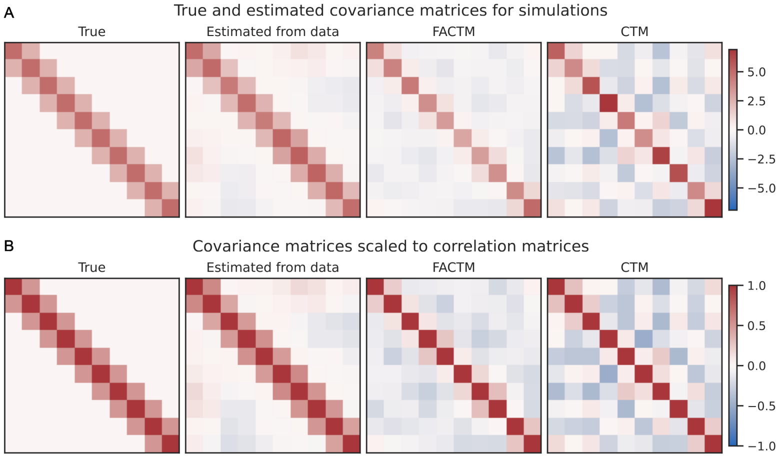

Figure B.5: Comparison of true ($\Sigma^{(0)}$) and estimated covariance matrices (**A**) and covariance matrices scaled to have 1 on the diagonal (**B**). From left to right: the true matrix used in the simulations; a matrix estimated using the true values of μ_n, η_n, and $\mu^{(0)}$; the matrix inferred by FACTM; and the matrix inferred by CTM.

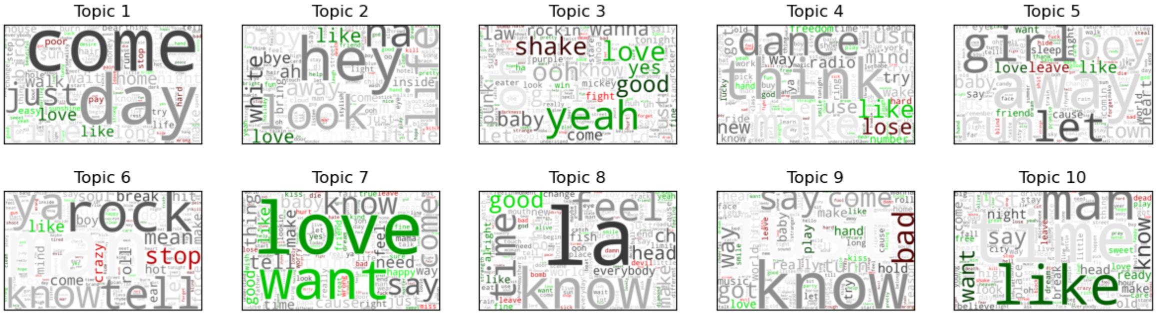

Figure B.6: Wordclouds representing the topics inferred by FACTM for the Mirex dataset. Words associated with positive sentiment are displayed in green, those with negative sentiment are shown in red, while the neutral words are gray. The size of each word corresponds to the probability of its occurrence within the topic.

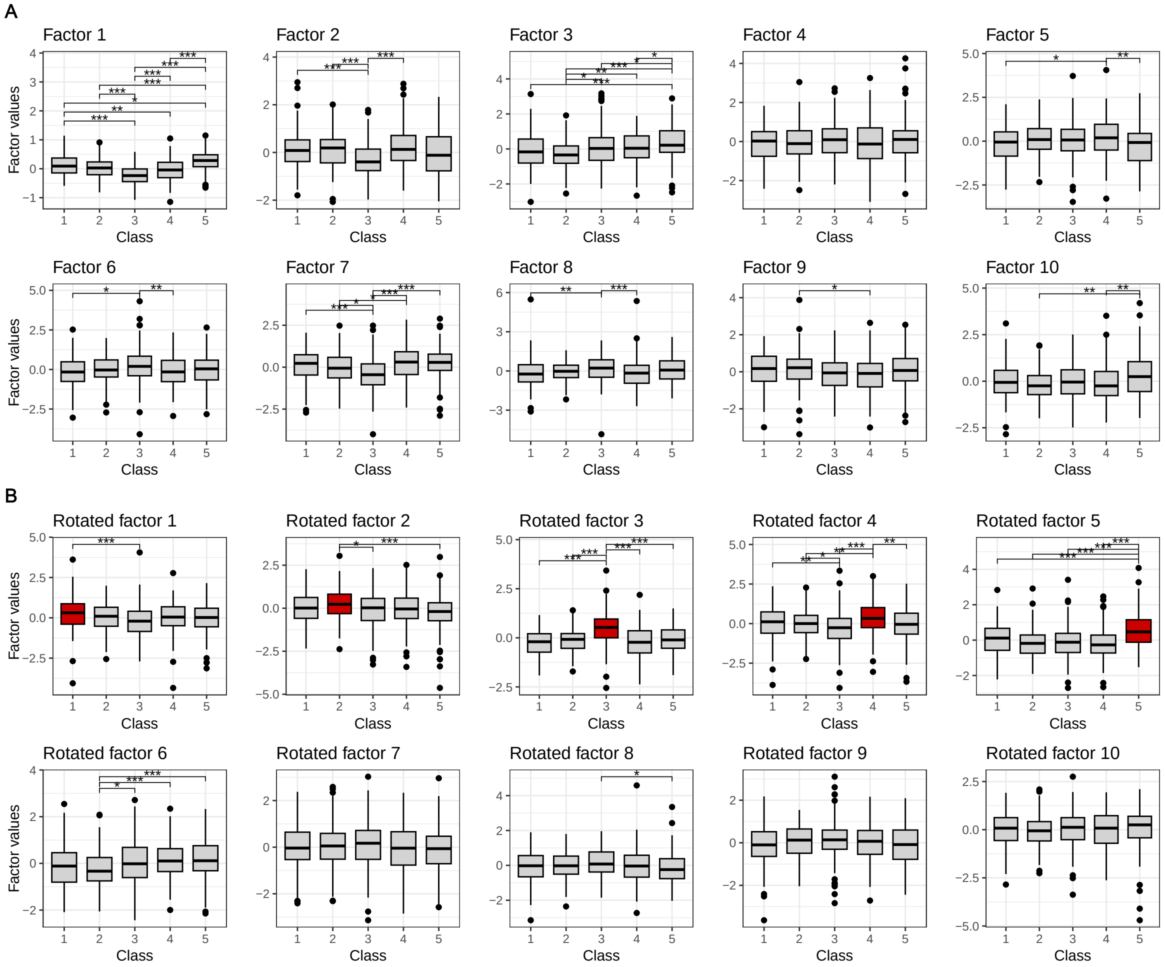

Figure B.7: The latent factors (**A**) and rotated latent factors (**B**) split by class membership of the samples. In red, we highlight the class and rotated factor pairs expected to be associated with each other by the construction of the rotation method. Significance of the differences between factor value distributions in the classes were evaluated with two-sided Wilcoxon pairwise tests. Stars denote significance of Bonferroni-adjusted p-values: $*** < 0.001$, $** < 0.01$, $* < 0.05$.

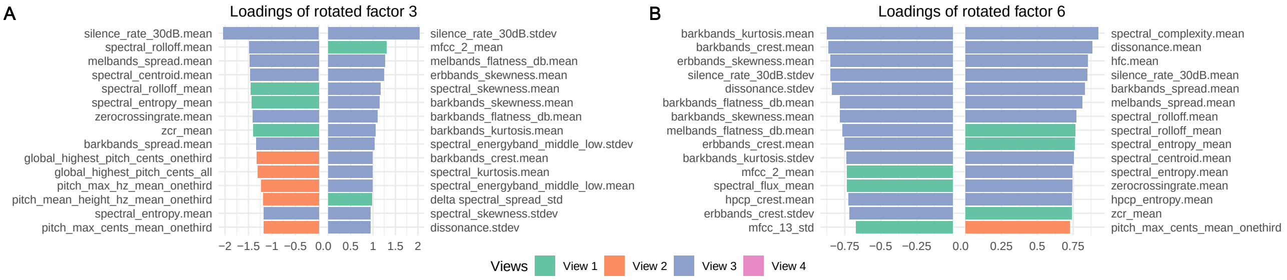

Figure B.8: Top-weighted features of simple views for rotated factor 3 (**A**) and rotated factor 6 (**B**), colored by the view to which they belong. Since the input data was quantile normalized, the loadings between views are comparable.

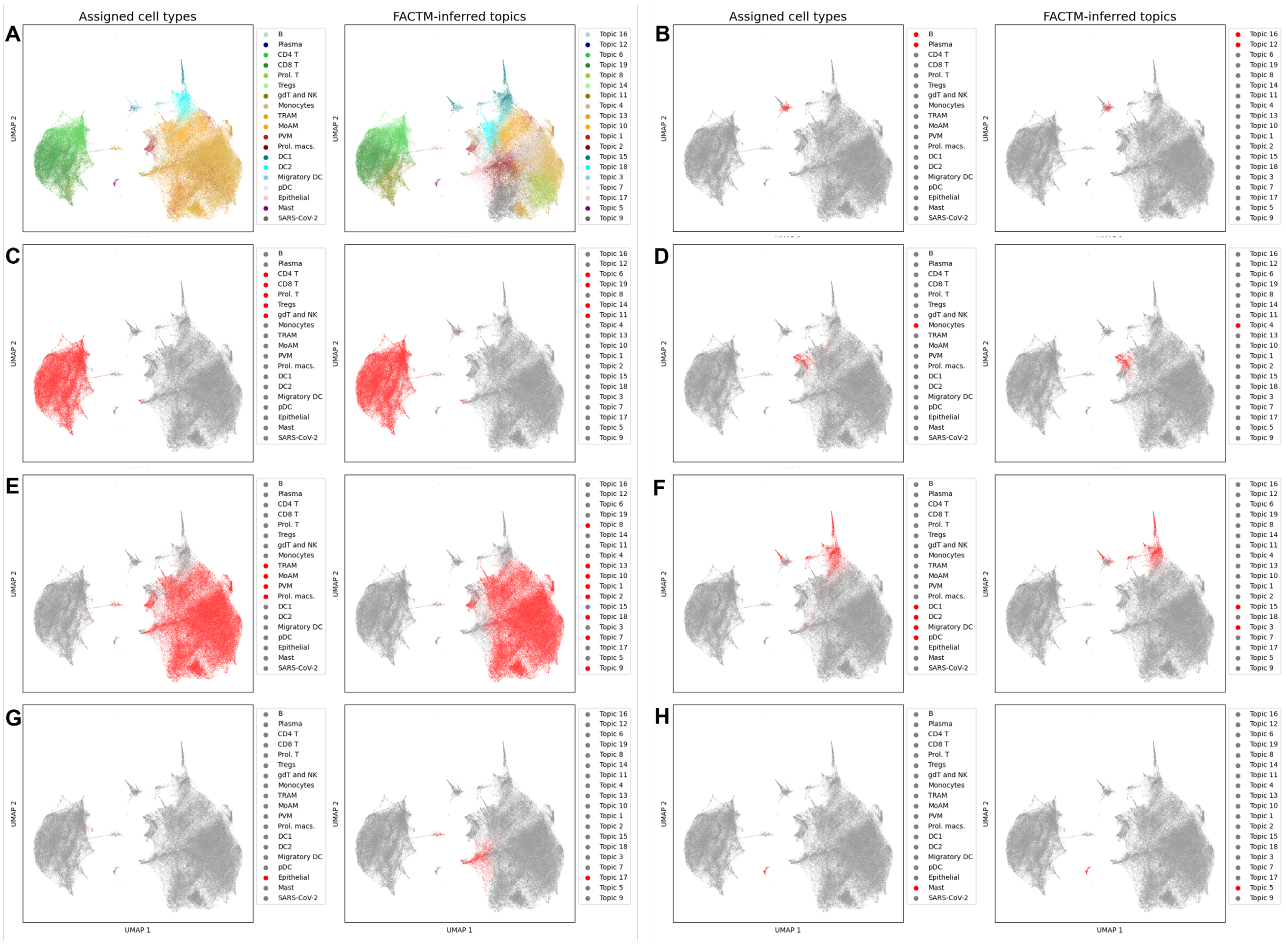

Figure B.9: UMAP plots comparing the clusters inferred by FACTM with the cell types assigned in the original source of the data. The UMAP representation is based on a matrix of 1,000 highly variable genes, with points colored according to the assigned clusters/cell types. Each point represents a single cell. **A.** Cell types and their corresponding topics, as in Fig. 7. **B. - H.** Comparison of more general cell types and the assigned clusters. Points are colored by selected cell types and topics (each topic is colored red once).

C ADDITIONAL TABLES

Table C.1: Supplement to Table 1 from the main text. Performance of the random forest classifier using inferred latent factors across the datasets for predicting assigned labels. Results are presented as the mean ROC-AUC and PR-AUC ± standard deviation over 10-fold cross-validation.

Measure	Model	Mirex	CMU-Mosei	CMU-Mosi
PR-AUC	FA+CTM	0.39 ± 0.04	0.70 ± 0.01	0.66 ± 0.05
	FACTM	0.39 ± 0.03	0.70 ± 0.01	$\mathbf{0.67} \pm 0.04$
	MOFA	0.39 ± 0.04	$\mathbf{0.71} \pm 0.01$	0.65 ± 0.05
	muVI	$\mathbf{0.40} \pm 0.03$	0.67 ± 0.01	$\mathbf{0.67} \pm 0.04$
ROC-AUC	FA+CTM	$\mathbf{0.68} \pm 0.04$	$\mathbf{0.73} \pm 0.01$	0.67 ± 0.04
	FACTM	0.67 ± 0.02	$\mathbf{0.73} \pm 0.01$	$\mathbf{0.68} \pm 0.04$
	MOFA	$\mathbf{0.68} \pm 0.03$	$\mathbf{0.73} \pm 0.01$	0.66 ± 0.04
	muVI	$\mathbf{0.68} \pm 0.02$	0.70 ± 0.01	0.67 ± 0.04

Table C.2: Supplement to Table 2 from the main text. Performance of the random forest classifier using model-specific observation-level representation from the structured part of data across the datasets. Results are presented as the mean ROC-AUC and PR-AUC ± standard deviation over 10-fold cross-validation.

Measure	Model	Mirex	CMU-Mosei	CMU-Mosi
PR-AUC	CTM	0.29 ± 0.03	0.52 ± 0.02	0.56 ± 0.05
	FACTM	$\mathbf{0.35} \pm 0.06$	0.60 ± 0.01	$\mathbf{0.60} \pm 0.05$
	LDA	0.33 ± 0.06	$\mathbf{0.63} \pm 0.02$	0.56 ± 0.04
ROC-AUC	CTM	0.57 ± 0.04	0.55 ± 0.02	0.57 ± 0.05
	FACTM	$\mathbf{0.64} \pm 0.06$	0.63 ± 0.01	$\mathbf{0.61} \pm 0.06$
	LDA	0.62 ± 0.05	$\mathbf{0.66} \pm 0.01$	0.57 ± 0.04

D NOTATION

Distributions We use the following notation to represent probability distributions:

- $\mathcal{N}(\mu, \sigma^2)$ - normal distribution (μ - mean, σ^2 - variance)
- $\mathcal{N}_L(\mu, \Sigma)$ - multivariate normal distribution (μ - mean vector, Σ - covariance matrix)
- $\text{Bern}(p)$ - Bernoulli distribution (p - success probability)
- $\mathcal{G}(a, b)$ - Gamma distribution (a - shape parameter, b - rate parameter)
- $\text{Beta}(a, b)$ - Beta distribution (a - shape parameter, b - shape parameter)
- $\text{Dir}(\alpha)$ - Dirichlet distribution (α - concentration parameter vector)
- $\text{Mult}(n, p)$ - Multinomial distribution (n - number of trials, p - probability vector)

FACTM notation Table D.3 provides an overview of the notation used in this article to describe the model.

Table D.3: Notation used in FACTM: Indices, observed variables, and latent variables.

Notation	**Meaning**	**Notes**
N	Number of samples (observations)	
M	Number of single views (modalities)	
D^m	Number of features in modality m	Can vary across modalities
K	Number of latent factors	Hyperparameter
L	Number of clusters (topics)	Hyperparameter
I_n	Number of data points in sample n	Can vary across samples
J_i	Number of observed objects in data point i	Can vary across data points
G	Number of distinct object types (e.g. distinct words)	
$y^m_{n,d}$	Feature d for sample n in modality m	Observed
$\overline{y}_{n,i,j}$	Object j in data point i of sample n $\overline{y}_{n,i,j} \in \{1, 2, \dots, G\}$	Observed
$\overline{y}_{n,i,g}$	Number of objects of type g in data point i of sample n $\overline{y}_{n,i,g} = \sum_{j=1}^{J_i} \mathbb{I}(\overline{y}_{n,i,j} = g)$	Observed
$\{\overline{y}_{n,i,g}\}_{g=1}^{G}$	Object counts by type in data point i of sample n	Observed
$z_{n,k}$	Latent factor k for sample n	Latent factors
$w^m_{d,k}$	Loading of factor k for feature d in modality m $w^m_{d,k} = \tilde{w}^m_{d,k} s^m_{d,k}$	Loadings for simple views
$\tilde{w}^m_{d,k}$	Normal component of loading $w^m_{d,k}$	Loadings - ARD prior
α^m_k	Precision parameter for $\tilde{w}^m_{\cdot,k}$ Hyperparameters: a_0^α, b_0^α	Loadings - ARD prior
$s^m_{d,k}$	Binary component of loading $w^m_{d,k}$	Loadings - spike-and-slab prior
θ^m_k	Probability of success for $s^m_{\cdot,k}$ Hyperparameter: a_0^θ, b_0^θ	Loadings - spike-and-slab prior
$\overline{w}_{l,k}$	Loading of factor k for cluster (topic) l	Loadings for structured view
$\overline{\alpha}_k$	Precision parameter for $\overline{w}_{\cdot,k}$ Hyperparameters: $\overline{a}_0^\alpha$, $\overline{b}_0^\alpha$	Loadings - ARD prior
$\mu_{n,l}$	Modification of population-level mean $\mu^{(0)}$ for sample n and cluster l (sample-specific modification of cluster proportions) Hyperparameter: t	
$\mu^{(0)}$, $\Sigma^{(0)}$	Population-level mean and covariance	Dimensions: L and $L \times L$
η_n	After softmax transformation, cluster distribution for sample n	Dimension: L
$\xi_{n,i}$	Cluster assignment for data point i in sample n	
β_l	Object type (word) distribution for cluster (topic) l Hyperparameter: α_0^β	Dimension: G