Sparse Maximum Margin Learning from Multimodal Human Behavioral Patterns

Ervine Zheng, Qi Yu, Zhi Zheng

Rochester Institute of Technology
{mxz5733, qi.yu, zhzbme}@rit.edu

Abstract

We propose a multimodal data fusion framework to systematically analyze human behavioral data from specialized domains that are inherently dynamic, sparse, and heterogeneous. We develop a two-tier architecture of probabilistic mixtures, where the lower tier leverages parametric distributions from the exponential family to extract significant behavioral patterns from each data modality. These patterns are then organized into a dynamic latent state space at the higher tier to fuse patterns from different modalities. In addition, our framework jointly performs pattern discovery and maximum-margin learning for downstream classification tasks by using a group-wise sparse prior that regularizes the coefficients of the maximum-margin classifier. Therefore, the discovered patterns are highly interpretable and discriminative to support downstream classification tasks. Experiments on real-world behavioral data from medical and psychological domains demonstrate that our framework discovers meaningful multimodal behavioral patterns with improved interpretability and prediction performance.

Introduction

Analyzing human behavior is an important and broad research topic in various areas, including decision science, economics, sociology, and many more (Pantic et al. 2007). Human behavioral data usually involves multiple modalities (Barros et al. 2018; Rasouli, Kotseruba, and Tsotsos 2017), such as verbal communications, gestures, eye gazes, and facial expressions. The research of human behaviors has significantly benefited from the technological advances in multimodal data fusion (Song, Morency, and Davis 2012).

Data fusion can capture the complex relationship across modalities and provide predictive information for understanding the data. For example, factorization-based models decompose the data into the shared factor matrix and the matrix capturing the uniqueness of each modality (Correa et al. 2010; Sorber, Van Barel, and De Lathauwer 2015). Bayesian graphical models capture the joint probability of multiple modalities, or conditional probability of cross-modal relations (Velivelli and Huang 2008; Nakamura et al. 2011). In recent years, deep neural networks (DNN) have been developed for data fusion, aiming to model complicated relationships among modalities through multi-level feature extraction and integration (Hori et al. 2017; Zhou et al. 2019), reduce noise through a multi-feature autoencoder (Ma et al. 2016), and effectively handle missing data through a generative variational module (Seo et al. 2018). With a deep network, multimodal feature vectors can be concatenated for fusion purposes (Baltrušaitis, Ahuja, and Morency 2018).

However, applying data fusion models to human behavioral studies still faces fundamental challenges in specialized domains, such as psychology and health. One primary reason is that data collection is usually based on rigorously designed experiments involving human subjects, which is difficult to conduct on a large scale. The limited behavioral data may hinder the application of the existing data-driven models that require massive training data (Wu and Goodman 2018; Tsai et al. 2019; Kumar et al. 2021). Besides data scarcity, there are several critical requirements of behavioral analytics that may not be simultaneously satisfied by existing data-driven models: 1) Human behavioral data is inherently dynamic and multimodal, while behavioral research in psychology and health domains usually requires discovering interpretable patterns from complex behavior. 2) Aside from pattern discovery, research in those specialized domains may involve classification tasks, such as predicting which social group a person belongs to based on the observed behaviors. However, if pattern discovery (unsupervised) and classification (supervised) are conducted in isolation, some discovered patterns may not contribute to classification and may cause model overfitting. 3) Some modalities of human behavioral data may be highly noisy or exhibit no easily distinguishable patterns. Directly using such data for downstream domain research without systematic feature selection will negatively impact the final decision-making process.

To address the fundamental challenges and meet all critical requirements simultaneously, we propose Sparse Maximum Margin learning from Multimodal Recurrent States (SM^2-MRS) of human behavioral patterns. The MRS model extracts significant patterns from the observed human behavioral data and uses a latent state space to dynamically fuse the patterns from different modalities. In particular, MRS formulates a two-tier probabilistic mixture model. The lower tier models the observations from each modality as a mixture of component distributions from the exponential family (EF). Mixtures of EF distributions are adopted due to the following reasons. First, most behavioral data are low-dimensional in

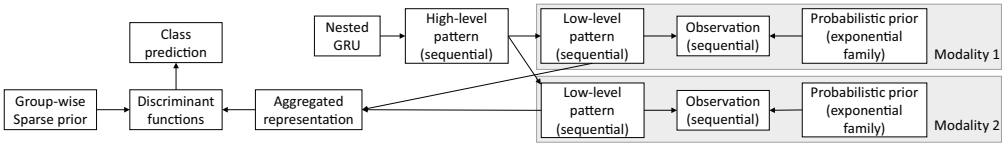


Figure 1: Overall architecture. For modeling human behaviors, our framework uses mixture of patterns to discover interpretable behavioral patterns from each modality; nested GRU to capture temporal dependency; max-margin classifier to learn a robust decision boundary and classify sequential behavioral data; sparse max-margin learning to identify important patterns to avoid overfitting. Please refer to the Experiment section for use cases.

their original forms (*e.g.,* eye gazes and motions) or with a higher dimension but very sparse (*e.g.,* verbal narrations with one-hot encoding). For the former, EF distributions are suitable to directly model low-dimensional data; for the latter, the parametric forms of EF usually lead to a good (approximate) distribution by learning only a few natural parameters. So it can balance the bias and variance to avoid overfitting on sparse data. Second, by linking the sufficient statistics of EF distributions and the observed data samples, the discovered mixture components (*i.e.,* low-level patterns) can be conveniently interpreted.

To fuse multimodal behavioral data, we further introduce high-level patterns, which are a mixture of low-level patterns. Each high-level pattern can be explained using a subset of representative low-level patterns from different modalities. Furthermore, we build a nested gated recurrent unit (GRU) to organize the patterns into a state space, and capture the temporal dynamics in human behaviors.

To support downstream classification tasks and decision-making, we propose to jointly perform dynamic data fusion and sparse maximum-margin learning (SM^2) from the space of multimodal behavioral patterns. The joint learning aims to extract interpretable patterns that can explain human behaviors and have sufficient discriminative power to support classification. We ensure the sparsity of useful patterns using a unique hierarchical group-wise Laplace prior. The inference process is developed to tackle the complex (non-conjugate) interactions between the sparse Laplace prior of the SM^2 coefficients and the latent variables in the MRS. As part of the inference process, we derive a key property of a Laplace distribution (see Theorem 1). It ensures an analytical form of the (approximate) posterior of the SM^2 coefficients, which allows us to solve a constrained quadratic dual problem (see Theorem 2) for efficient posterior inference.

Figure 1 shows the overall architecture of the proposed SM^2-MRS. Our main contributions include:

- a dynamic multimodal fusion framework to jointly perform data fusion and maximum-margin learning from human behavioral data in specialized domains,
- a two-tier probabilistic mixture model that leverages EF distributions to discover low-level behavioral patterns, which are then dynamically fused at the high level to accommodate data sparsity, heterogeneity, and temporal dynamics simultaneously,
- a group-wise sparse prior that automatically selects discriminative patterns with improved interpretability and reduced risk of overfitting,
- an efficient algorithm to solve the max-margin posterior inference via quadratic programming with a theoretical guarantee.

We conduct experiments on two real-world datasets from medical and psychological domains. Our model extracts interpretable multimodal behavioral patterns and provides insights to benefit the research in those domains. Meanwhile, the proposed model also achieves the best prediction performance on supervised learning tasks, which indicates that it has the potential to assist domain experts in human-machine collaborative decision-making.

Related Works

Data fusion. Data fusion models leverage multiple modalities of the observed data to capture complementary information and make robust predictions. For probabilistic models, matrix decomposition methods factorize the data into a matrix capturing shared factors and a matrix capturing the uniqueness (Sorber, Van Barel, and De Lathauwer 2015). However, the inferred feature representations are usually continuous vectors that may not have an understandable interpretation.Bayesian graphical models fuse data by using a joint pattern to govern multiple modalities (Nakamura et al. 2011). However, those models may end up with too many patterns, especially when data modalities are just loosely coupled. For deep learning models, deep belief networks transfer the representations from individual modalities into the semantic features in the shared space (Srivastava and Salakhutdinov 2012; Al-Waisy et al. 2018). Autoencoder-based models use the encoder-decoder architecture to extract modality-specific representations and shared representations (Wu and Goodman 2018; Tsai et al. 2019; Kumar et al. 2021). Recurrent neural network-based models capture the temporal dependency from sequential data by fusing recurrent hidden units (Tsai et al. 2019; Sano et al. 2018). In summary, DNN models capture cross-modal interaction through joint representation learning. However, for many specialized domains such as psychology and health, limited annotated training data is a common issue, and DNNs may suffer from deteriorated performance due to overfitting. In addition, DNNs may not be easily interpretable (Gao et al. 2020). Existing methods provide interpretability by introducing feature importance or attention weights (Hsu, Zhang, and Glass 2017; Heo et al. 2018). However, they are based on latent representations with high-dimensional vectors, which are still difficult to interpret due to a lack of semantic meanings. In contrast, our model uses EF distributions to model behavioral patterns, and those

patterns can be visualized and interpreted by domain experts.

Maximum margin models. Maximum margin learning finds a hyperplane to classify data points so that the distance to the nearest data point on each side is maximized. It can be integrated with probabilistic pattern discovery to train Markov networks (Zhu and Xing 2009). Extensions include single-modal data analysis on text analytics, image understanding, and representation learning (Zhu, Ahmed, and Xing 2012; Wang and Mori 2011; Tu et al. 2016), and multiview subspace learning that assumes weak conditional independence among heterogeneous observations (Chen et al. 2012). However, existing models may not work well on behavioral data with complex cross-modal interactions. In addition, the maximum margin classifier for multi-class classification requires training several scoring functions, one for each class. Conventional feature selection techniques such as vanilla l_1-regularization and sparse priors do not work well, because removing a feature requires its corresponding coefficients in all scoring functions shrunk to 0. In contrast, our group-wise sparse prior addresses the above issues.

The Proposed Framework

We aim to discover informative and interpretable patterns from human behavioral data, and leverage the patterns to support downstream classification. In specialized domains, behavioral data is usually collected from experiments with a group of participants. The data is usually multimodal and sequential. We first introduce the concepts.

- *Modality*: Behavioral data is usually in multiple heterogeneous types (*e.g.,* language and action). Each type is considered a modality, indexed by m, where $1 \leq m \leq M$.
- *Data instance*: In experiments, participants may be instructed to perform different tasks. For example, in an experiment that studies children's behavior, a group of children is invited to play several specifically designed video games. The behavioral data collected from one participant (*i.e.,* a child) and one task (*i.e.,* a game) is considered a data instance, indexed by l, where $1 \leq l \leq L$.
- *Time step*: Behavioral data is usually temporal. We partition the timeline of each data instance into multiple slices with equal duration, each treated as a time step, indexed by t, where $1 \leq t \leq T$.
- *Observation*: In each time step, we have one observation for each modality of the behavioral data. We denote the observation for data instance l modality m at time t as $x_t^{(l,m)}$, and the whole sequence of data as $X^{(l,m)} = \left\{x_1^{(l,m)}, ..., x_t^{(l,m)}, ..., x_T^{(l,m)}\right\}$.
- *Class label*: Each data instance corresponds to one class label that describes the nature of the data instance. In the previous example (behavioral experiment), the class label could be the psychological group of the child. We denote the label for data instance l as $y^{(l)}$.

Data Fusion Using Multimodal Behavioral Patterns

Following the notations introduced in previous section, we consider a dataset with M modalities and L data instances. Each data instance has T observations and one corresponding response variable, *e.g.,* the class label. We assume the observations are generated from the exponential family (EF):

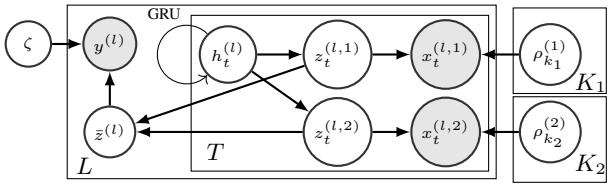


Figure 2: Graphical illustration of the framework. Observed data instances, behavioral patterns, and high-level patterns, are denoted by x, z, and h, respectively. Data instances and time steps are indexed by l and t, respectively. Assume data instance l has two modalities (denoted by $x^{(l,1)}, x^{(l,2)}$). Class labels are denoted by y. Other variables are introduced along with the prior distributions (see next sections for details)

$$\begin{aligned} &p(x_t^{(l,m)}|z_t^{(l,m)} = k, \rho_k^{(m)}) \\ &= h(x_t^{(l,m)})g(\rho_k^{(m)}) \exp((\rho_k^{(m)})^\top \tau(x_t^{(l,m)})) \end{aligned} \tag{1}$$

where $z_t^{(l,m)}$ is low-level pattern assignment, $x_t^{(l,m)}$ is observation. $\rho_k^{(m)}$ is the natural parameter for pattern k in modality m. $h(\cdot)$, $g(\cdot)$ and $\tau(\cdot)$ are known functions corresponding to a specific distribution (*e.g.,* Gaussian or Multinomial). We place a conjugate prior on each pattern k as:

$$\begin{aligned} &p(\rho_k^{(m)}|\chi_k^{(m)}, v_k^{(m)}) \\ &\propto g(\rho_k^{(m)})^{v_k^{(m)}} \exp(v_k^{(m)}(\rho_k^{(m)})^\top \chi_k^{(m)}) \end{aligned} \tag{2}$$

where $\chi_k^{(m)}$ and $v_k^{(m)}$ are the parameters of the conjugate prior. Here we use a general form of EF distributions, while the specific forms (*e.g.,* Gaussian or Dirichlet) depend on the nature of the data and prior domain knowledge. Please refer to the experiment section for some examples.

For each timestep, we further introduce a high-level pattern that describes the joint distribution of low-level patterns from each modality. The conditional dependency between low-level pattern $z_t^{(l,m)}$ and high-level pattern $h_t^{(l)}$ is modeled with a categorical distribution parameterized by $\theta^{(m)}$.

$$p(z_t^{(l,m)}|h_t^{(l)} = j) \sim \text{Cat}(\theta_j^{(m)}), \quad \theta_j^{(m)} \sim \text{Dir}(o_0) \tag{3}$$

where $h_t^{(l)}$ is high-level pattern assignment for data instance l at timestep t, j is the index of high-level pattern. $\text{Dir}(o_0)$ is a global Dirichlet prior.

We use a nested single-layer gated recurrent unit (GRU) to model the transition of high-level patterns over time. GRU is a variant of the widely-used long-short term memory network (LSTM) for modeling temporal data, while it lacks an output gate and thus has fewer parameters.

$$p(h_t^{(l)}|h_{1:t-1}^{(l)}) \sim \text{Cat}(\text{GRU}(h_{1:t-1}^{(l)})) \tag{4}$$

The output of GRU (after softmax transformation) is a probability vector, which is considered the parameter of categorical distribution to generate the high-level pattern sequence.

Therefore, when considered as a whole, the single-layer GRU has a probabilistic interpretation as specified in (4).

The joint distribution of patterns and observations is:

$$\begin{aligned} & p(X, Z, H, \rho, \theta) \\ = & \prod_{m,k} p(\rho_k^{(m)}) \prod\nolimits_{m,j} p(\theta_j^{(m)}) \prod\nolimits_{l,t} p(h_t^{(l)} | h_{1:t-1}^{(l)}) \\ & \prod\nolimits_{l,m,t} p(z_t^{(l,m)} | h_t^{(l)}, \theta^{(m)}) \prod\nolimits_{l,m,t} p(x_t^{(l,m)} | z_t^{(l,m)}, \rho) \end{aligned} \tag{5}$$

Integrating Maximum-Margin Learning and Group-Wise Regularization

The latent multimodal patterns are useful for downstream classification. We propose using maximum-margin learning for classification. Consider R classes and denote the class variable for each data instance as $y^{(l)} \in \{1...R\}$. For a class y, We define a linear discriminant function F: $F(y, z^{(l)}) = \zeta_y^\top \bar{z}^{(l)}$, where $\bar{z}^{(l)} = 1/T(\sum_t z_t^{(l,1)}, ..., \sum_t z_t^{(l,m)})^\top$ is the aggregated latent pattern assignments of data instance l's observations, and $\bar{z}^{(l)}$ can be considered the representation of data instance l. ζ_y is a class-specific coefficient vector associated with class y in the linear discriminant functions. Given a data instance l, the discriminant functions calculate a score for each class, and the ground-truth class's score is encouraged to be greater than any other class's score by a certain margin (empirically set to 1):

$$\forall y \neq y^{(l)} : F(y^{(l)}, z^{(l)}) - F(y, z^{(l)}) > 1 \tag{6}$$

Maximum margin learning has its foundation in support vector machines. Compared with alternative design choices for classification (*e.g.,* logistic regression), the benefit of integrating maximum margin learning with probabilistic models is to learn an effective decision boundary with an improved generalization capability (Zhu, Ahmed, and Xing 2012). Our experiments also show that maximum margin learning achieves better results than alternative designs.

Assume the length of ζ_y is K, which equals to the number of total latent patterns. For example, in a data set with 2 modalities, if the number of patterns of each modality is 2 and 3, then $K = 2 + 3 = 5$. Let ζ denote a vector of concatenating ζ_y over all R classes, and let $f(y, \bar{z}^{(l)})$ denote a vector whose $(y-1)K+1$ to yK entries are from $\bar{z}^{(l)}$ and all others are 0. Then F can be re-written as

$$F(y, z^{(l)}) = \zeta^\top f(y, \bar{z}^{(l)}) \tag{7}$$

In addition, we introduce a group-wise sparse prior to identify informative patterns that contribute to the classification while discarding non-informative patterns. Those informative patterns can be used by experts for further domain-specific analysis. Specifically, we enforce the sparsity on coefficients ζ's through a group-wise Laplace prior with Gaussian-exponential hierarchical representation

$$p(\zeta | s) = \mathcal{N}(\zeta | 0, \mathrm{diag}(s)), p(s_k) = \frac{\lambda}{2} \exp\left(-\frac{\lambda}{2} s_k\right) \tag{8}$$

where $s > 0$ is a vector and $s_{(y-1)K+k} = s_k$ for any $y \in \{1...R\}$. The variance s_k of latent pattern k's coefficients is shared across the discriminant functions of all classes. With the group-wise Laplace prior, some s_k may be driven to 0 during model training, forcing the corresponding ζ's to approach 0 and thus deleting those latent patterns for classification. We also provide an intuitive explanation of maximum margin learning and group-wise regularization in the Appendix.

Posterior Inference

Exact posterior inference is intractable due to the dependency among the following parts: (I) parameters ρ and θ, and assignments Z and H, (II) the nested GRU, (III) the maximum margin coefficient ζ and parameter s in the hierarchical group-wise Laplace prior. However, the inference procedure can be formulated as variational inference, which optimizes each variational distribution iteratively (Bishop and Nasrabadi 2006). To deal with part (I), we define a variational distribution $q(Z, H, \theta, \rho)$ that can be factorized as:

$$\begin{aligned} & q(Z, H, \theta, \rho) = \prod_{m,j} q(\theta_j^{(m)} | \omega_j^{(m)}) \prod_{m,k} q(\rho_k^{(m)} | \xi_k^{(m)}) \\ & \prod_{l,t} q(h_t^{(l)} | \gamma_t^{(l)}) \prod_{l,m,t} q(z_t^{(l,m)} | \phi_t^{(l)}) \end{aligned} \tag{9}$$

where ω, γ, ξ and ϕ are the parameters of variational distributions. An optimal $q(Z, H, \theta, \rho)$ can be obtained by minimizing the KL divergence $KL(q \| p)$ between the true posterior and variational approximation, which is equivalent to maximizing the *evidence lower bound* $L[q(Z, H, \theta, \rho)]$:

$$\begin{aligned} & L[q(Z, H, \theta, \rho)] = \int\int q(Z, H, \theta, \rho)(\ln p(X, Z, H, \theta, \rho) \\ & - \ln q(Z, H, \theta, \rho))\mathrm{d}Z\mathrm{d}H\mathrm{d}\theta\mathrm{d}\rho. \end{aligned}$$

Inference for **part (I)** can be handled using standard variational inference because of the conjugacy between variables, ensured by the design of the framework. Due to the space limit, we provide the detailed updates rule for ω, γ, ξ, ϕ in the Appendix. For **part (II)**, we note that given the assignments Z, the nested GRU is independent from other parts and its parameters can be updated using stochastic gradient descent (details are in the Appendix). For **part (III)**, we need to infer the posterior distribution of ζ and s. We consider the marginal distribution $p(\zeta) = \int p(\zeta | s) p(s | \lambda) \mathrm{d}s$, where $p(s|\lambda) = \prod_k p(s_k|\lambda)$ and $p(\zeta|s) = \prod_{k,y} p(\zeta_{(y-1)K+k} | s_k)$. Assume $q(\zeta, s) = q(\zeta)q(s)$ and apply Jensen's inequality, we get the *evidence lower bound* $L[q(\zeta, s)]$ as

$$\begin{aligned} & KL(q(\zeta) \| p(\zeta)) \\ & \leq -H(q(\zeta)) - \iint q(\zeta) q(s) \ln \frac{p(\zeta | s) p(s | \lambda)}{q(s)} \mathrm{d}s\mathrm{d}\zeta \\ & = L[q(\zeta, s)] \end{aligned} \tag{10}$$

Inference of ζ and s needs to deal with the unique challenges due to the maximum margin constraints and the group-wise sparse prior. The major results are summarized below:

The (approximate) posterior inference of maximum-margin multimodal data fusion with group-wise Laplace regularization is achieved by solving:

$$\min_{q(Z,H,\theta,\rho),q(\zeta,s),\epsilon^{(l)}} -L[q(Z,H,\theta,\rho)] - L[q(\zeta,s)] + C\sum\nolimits_l \epsilon^{(l)}$$
$$\text{s.t.} \quad \forall l, y \neq y^{(l)}: \quad \mathbb{E}_q[\zeta^\top \triangle f^{(l)}(y)] \geq 1 - \epsilon^{(l)}, \quad \epsilon^{(l)} \geq 0 \tag{11}$$

where $\triangle f^{(l)}(y) = f(y^{(l)}, \bar{z}^{(l)}) - f(y \neq y^{(l)}, \bar{z}^{(l)})$, and $\epsilon^{(l)}$ is the slack variable for instance l. The first two terms of the objective function formulates the evidence lower bound for variational approximation, while the third term is the soft-margin penalty.

Notice that $q(Z, H, \theta, \rho)$ is irrelevant with the constraint and is solved in part (I). Now we calculate $q(\zeta)$ and $q(s)$. In particular, variational distribution $q(\zeta)$ is assumed a conjugate Gaussian distribution $q(\zeta) \sim \mathcal{N}(\zeta|\mu_\zeta, \Sigma_\zeta)$:

$$\mu_\zeta = \Sigma_\zeta \sum_{l, y\neq y^{(l)}} v^{(l)}(y) \triangle f^{(l)}(y), \quad \Sigma_\zeta = \left(\text{diag}(\mathbb{E}_q[s^{-1}])\right)^{-1} \tag{12}$$

The $v^{(l)}$'s in (12) are Lagrangian multipliers introduced to handle the constraints in (11). In addition, computing Σ_ζ requires to calculate $\mathbb{E}_q[s^{-1}]$. Let s_k denote k-th element of s. To calculate $v^{(l)}$ and $\mathbb{E}_q[s_k^{-1}]$, we introduce the following theorems (Detailed proofs are provided in the Appendix).

Theorem 1. *The variational distribution of $q(s_k)$ takes the following form:*

$$q(s_k) \propto \exp\left(-\frac{1}{2}\lambda s_k\right) \prod\nolimits_y \mathcal{N}\left(\sqrt{\mathbb{E}_q(\zeta_{(y-1)K+k})^2}|0, s_k\right) \tag{13}$$

In particular, the expectation of s_k^{-1} is given by

$$\mathbb{E}_q(\frac{1}{s_k}) = \frac{\lambda}{d} \sum_{r=0}^{R_0+1} \frac{(R_0+r+1)!}{r!(R_0+1-r)!(2d)^r} / \sum_{r=0}^{R_0} \frac{(R_0+r)!}{r!(R_0-r)!(2d)^r} \tag{14}$$

where $d = \sqrt{\lambda \sum_y \mathbb{E}(\zeta_{(y-1)K+k})^2}$, and $R_0 = \frac{R}{2} - \frac{1}{2}$.

Theorem 2. *The Lagrangian multipliers $v^{(l)}$'s of the primal problem* (11) *can be computed by solving a dual problem:*

$$\max_v -\frac{1}{2}\eta^\top \Sigma_\zeta \eta + \sum_l \sum_{y\neq y^{(l)}} v^{(l)}(y),$$
$$\textit{s.t.} \sum\nolimits_{y\neq y^{(l)}} v^{(l)}(y) \in [0, C], \quad v^{(l)}(y) \geq 0 \tag{15}$$

where $\eta = \sum_{l, y\neq y^{(l)}} v^{(l)}(y) \triangle f^{(l)}(y)$.

The dual problem is a constrained quadratic problem, which can be solved using standard QP solvers. Once $v^{(l)}$'s are computed, they can be plugged in (12) along with $\mathbb{E}_q(s_k^{-1})$ given by (14) to evaluate $q(\zeta)$. The detailed inference process is summarized in the Appendix.

Once the model is trained, given some new data, the model can predict the corresponding class labels and discover significant latent patterns to support domain research.

Experiments

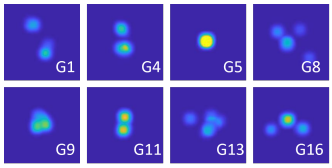
We present our experimental results on two real-world behavioral studies. We also collaborated with domain experts to interpret our findings. The collection and usage of data have received rigorous Institutional Review Board review. The appendix and source code are presented in (Zheng, Yu, and Zheng 2023).

Dataset I: Behavioral Study in Medical Domain

We conducted experiments where physicians work towards a diagnosis by viewing medical images and describing the image content and their analysis. The first modality is verbal narrations recorded and transcribed into word tokens and timestamps. The second modality is physicians' eye gazes recorded by eye-tracking. Fixations (the gazes maintained on a location) are recorded in terms of location and duration. A total of 1,614 data instances (narrations and eye gaze sequences) were collected. There are 4 classes (*i.e.,* solitary, symmetry, multiple morphologies, and high-density lesions), which are annotated by a group of domain experts based on disease morphology. Our model aims to discover latent patterns and perform classification.

For pattern discovery of verbal narrations (first modality), we apply the Dirichlet-Categorical conjugate distributions from the EF family, where each narration is considered a mixture of topics. The setting is widely used in topic models (Blei, Ng, and Jordan 2003). Denote the word distribution for topic k as β_k. For one narration l, given the topic assignment at time t as $z_t^{(l,1)}$, the corresponding words are drawn from $w_t^{(l)} \sim \text{Cat}(\beta_{z_t^{(l,1)}})$. For eye gazes (second modality), we preprocess the data and generate a series of 2D heatmaps to describe the visually attended areas within a short period (Lin et al. 2013). Then all heatmaps are reshaped as vectors $a_t^{(l)}$. Since heatmaps take continuous values, a common choice is to use Gaussian distribution. For pattern c, assume its mean μ_c has a Gaussian prior $\mu_c \sim \mathcal{N}(0, \sigma_0 I)$, where I denotes identity matrix. Given the pattern assignment $z_t^{(l,2)}$, the corresponding heatmap is drawn from $a_t^{(l)} \sim \mathcal{N}(\mu_{z_t^{(l,2)}}, sI)$ (see the Appendix for details). The occurrences of latent patterns from both modalities are aggregated and used to perform joint maximum margin learning for classification.

We choose representative baselines based on their prediction performance and applicability to the datasets. The baselines are from three different categories, including *1) Single-modal baselines that are relevant to some individual components in our data fusion framework*: supervised LDA (Mcauliffe and Blei 2008) (sLDA) applied to narration, Hidden Markov Topic Model with logistic regression (HMTM) (Andrews and Vigliocco 2010) applied to narration, Maximum entropy discriminant models (Zhu, Ahmed, and Xing 2012) applied to the narration and gaze data respectively (MED-narr and MED-gaze). *2) Probabilistic multimodal baselines with competitive performance*: LDA-based Multimodal Categorization (LDAM) (Nakamura et al. 2011), Large-margin latent subspace learning (LLSL) (Chen et al. 2012). *3) Recent deep-learning based multimodal models with competitive performance*: Multimodal generative models



Topic	Informative Words
T2	on lupus nose areas face young
T3	nevus cell melanoma papule basal
T4	versus final over two certain
T8	diagnosis differential certainty include
T7	i'm think about for there's
T17	looks like also but likely

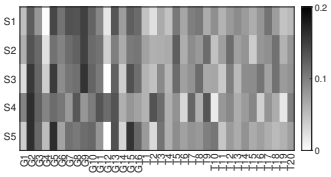


Figure 3: Visualization of some gaze patterns (left), inferred topics (middle), and high-level patterns (right).

Method	Accuracy	Method	Accuracy
Proposed	**85.5 ± 2.8**	LDAM	80.1 ± 3.7
sLDA	72.7 ± 3.9	LLSL	79.8 ± 3.5
HMTM	60.4 ± 4.1	AF	77.2 ± 3.7
MED-narr	73.7 ± 3.4	MGM	77.4 ± 4.2
MED-gaze	77.8 ± 2.9	FMR	76.9 ± 4.5

Table 1: Performance on Dataset I (Accuracy%)

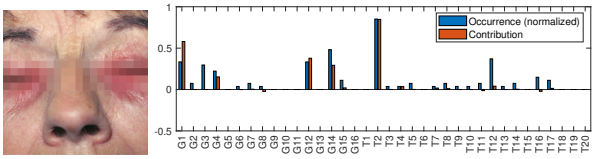


Figure 4: Case study of one physician making analysis on a medical image (left), the occurrence of gaze patterns G1-G16 and topics T1-T20 and their contribution (calculated in Eq (7)) to classification (right). It shows Switching (G1,G14) are the most prevalent gaze pattern and the description of facial area (T2) is the most prevalent topic. G1 and T2 contribute to classifying disease the most.

(MGM) (Wu and Goodman 2018) which applies variational autoencoder for weakly-supervised multimodal inference, Factorized multimodal representations (FMR) (Tsai et al. 2019) which integrates an LSTM with an encoder-decoder architecture, and AutoFuse (AF) (Kumar et al. 2021) which jointly optimize multimodal autoencoders and the classification layer. According to survey (Baltrušaitis, Ahuja, and Morency 2019), deep learning received more attention than probabilistic models in recent years. Therefore, probabilistic baselines are relatively older than deep learning baselines.

We report models' performance on classifying the disease morphology in Table 1. Our framework achieves the highest accuracy. Possible reasons are: single-modal baselines do not leverage cross-modal interaction; classical multimodal baselines are not customized to the domain requirements. Besides, their model architecture may be overly simplified; deep learning baselines are prone to overfitting; The behavioral dataset is on a small scale because data collection is costly. Our model can properly accommodate the data sparsity to discover interpretable patterns and model their dynamic interactions. It also learns robust decision boundaries through maximum-margin learning.

An illustrative example in Figure 3 shows the inferred topics with the most informative words, and the gaze patterns that describe the eye gaze locations during diagnosis. Topics 4 and 8 are about the diagnostic decision; Topics 7 and 17 are about reasoning process; while Topics 2 to 3 mainly describe patient demographics, body location, lesion configurations, and distribution. Those gaze patterns can be roughly interpreted as concentration pattern (*e.g.,* G5) characterized by a small concentrated area in the heatmap, switching pattern (*e.g.,* G4, G11) characterized by a few concentrated areas, and clutter pattern (*e.g.,* G8, G13) characterized by scattered fixation locations. The inferred high-level patterns allow us to study the interaction between eye gazes and verbal narrations, aiming to gain more insight into humans' cognitive reasoning process when performing image analysis. We visualize the occurrence probabilities of the patterns from both modalities in each state and have some interesting and intuitive observations: 1) Diagnosis decisions (*e.g.,* T8) are usually associated with the concentration patterns (*e.g.,* G5) in state S3, implying that physicians gaze at a specific area of abnormalities when they are trying to make diagnostic decisions. 2) Descriptions (*e.g.,* T2) are usually associated with the switching (*e.g.,* G11) and clutter pattern (*e.g.,* G8) in state S4. This finding is consistent with the intuition that people look around at unfamiliar images when gathering information. Figure 4 shows a case study of one physician's behavior.

Dataset II: Behavioral Study in Psychology

The second dataset was from a behavioral experiment that studied sensory processing in children with and without Autism Spectrum Disorder (ASD) (Koirala et al. 2021). A virtual reality (VR) interactive gaming system was developed and evaluated as a tool to assess the sensory processing patterns in children with ASD through gaming behaviors in response to sensory stimuli embedded in the painting game. The experiment involves 12 children with ASD and 12 typically developing children (the controlled group) as players. The game consisted of 12 episodes. In each episode, a maze occupied the majority of the game scene. There were also 3D objects rotating in fixed positions as visual distractions. The player could move a painting ball in 3D using a haptic robot. When the ball was pushed against the frontal surface of the maze, the touched part of the maze would turn yellow. The goal of the game was to paint the entire maze as soon as possible. All children played 12 game episodes. Each episode played by each player was considered a data instance. There are 3 modalities: 1) children's gaze position on the screen surface; 2) the 3D trajectory of the painting balls in the virtual space; 3) Painting Movement in Depth (PMD) which records how hard a child press or lift the haptic robot. Due to the presence of different sensations, participants might operate the painting ball differently in response to the stimuli. Our

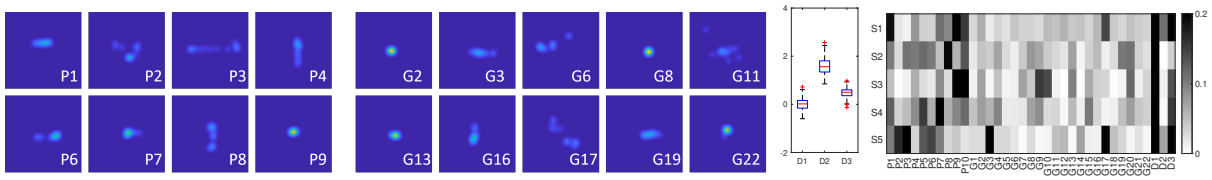


Figure 5: Visualization of some latent patterns from three modalities: Path (left), Gaze (middle-left) and PMD (middle-right), and the high-level patterns (right) .

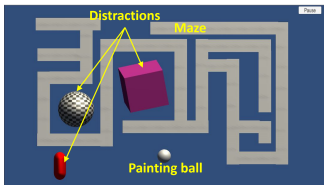


Figure 6: A game screen

Method	Accuracy	Method	Accuracy
Proposed	**74.2 ± 4.3**	LLSL	69.7 ± 4.5
MED-path	57.7 ± 5.8	AF	65.4 ± 4.1
MED-gaze	67.8 ± 4.4	MGM	67.1 ± 4.2
LDAM	68.9 ± 4.7	FMR	65.9 ± 5.0

Table 2: Performance on Dataset II (Accuracy%)

model aims to discover latent patterns from 3 modalities and perform classification of whether a child has ASD.

The ball's position was represented in the format of x and y coordinates of the game screen, and the eye gaze position was represented in the format of relative x and y coordinates with respect to the ball's position. Then, we aggregated the eye gaze and ball positions in each time slice, and preprocessed data in a similar way as introduced in the previous section to generate two heat maps, respectively. The heat maps contain rich information about how the participants controlled the ball in a short time period and how they visually attended to the ball. Our goal is to extract informative positions of gaze and ball on the screen as well as the ball's PMD patterns that disclose the behavioral difference between the players with ASD and the typically developing players. We used multivariate Gaussian priors to model the data generation process, and used the class label as additional supervised information to augment pattern discovery.

We report models' performance on classification in Table 2. The proposed framework outperforms baselines. It should be noted that classifying behavioral data is quite challenging, especially when the sample size is relatively small and the subjects' characteristics are heterogenous and noisy ($\sim$70% accuracy can be considered good (Cavallo et al. 2021)).

An illustrative example in Figure 5 shows the inferred patterns with the three modalities. Different patterns usually correspond to distinct behaviors. For instance, the path pattern P1 implies moving the ball slowly along the horizontal direction since the heat map reveals a highlighted horizontal line segment. Similarly, path pattern P2 implies moving the ball around a corner, and P3 implies moving the ball fast and horizontally. Gaze pattern G13 implies fixation on the ball most of the time with occasional deviation, as the heat map reveals a highlighted area at the center and some highlighted areas slightly away from the center. Similarly, gaze pattern G9 implies consistent fixation on the ball, and G17 implies frequent deviations from the ball. PMD pattern was visualized using boxplot. PMD pattern D1 indicates painting without lifting the ball, as the depth is around 0, while pattern D2 corresponds to lifting the ball far from the maze. We also study the relationship between the high- and low-level patterns from each modality by visualizing the normalized occurrence of patterns in each state, as shown in Figure 5. For instance, state (S1) is characterized by high occurrence of gaze pattern (G17) and path pattern (P1). We provide a **case study** of a player with ASD in the Appendix by analyzing behavioral patterns and their linkages to ASD symptoms.

Method	Dataset I	Dataset II
Proposed Design	**85.5 ± 2.8**	**74.2 ± 4.3**
Single-level Probabilistic Mixture	81.6 ± 3.2	71.9 ± 4.5
First-order Markov Structure	82.5 ± 3.1	70.1 ± 4.5
Multinomial regression	79.8 ± 3.5	68.1 ± 4.3
No regularization for sparsity	82.1 ± 2.9	72.0 ± 4.1
Vanilla L1 regularization	82.7 ± 2.5	72.5 ± 3.9

Table 3: Ablation Study on Two Datasets (Accuracy%)

Ablation study We evaluate alternative model design choices: 1) Single-level probabilistic mixture for pattern discovery, 2) First-order Markov structure for temporal dependency, 3) Multinomial regression for classification, 4) No regularization or vanilla L1 for sparsity. Details of ablation studies are provided in Appendix. We also evaluate the proposed group-wise regularization and L1 regularization in selecting discriminative patterns in the Appendix.

Conclusion

We propose a dynamic multimodal fusion framework to analyze human behavioral data from specialized domains. We design a two-tier probabilistic mixture model to discover interpretable behavioral patterns and dynamically fuse multimodal patterns. We develop maximum-margin learning with a group-wise sparse prior to select discriminative patterns for classification tasks. An efficient posterior inference algorithm is developed with theoretical proof. Our experiments on real-world human behavioral studies show promising results in pattern discovery and prediction accuracy, which demonstrate its great potential to facilitate human experts in critical decision-making and scientific discovery.

Acknowledgements

This research was partially supported by NSF IIS award IIS-1814450 and ONR award N00014-18-1-2875. Any opinions, findings, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the official views of any funding agency. We would like to thank the anonymous reviewers for reviewing the manuscript.

References

Al-Waisy, A. S.; Qahwaji, R.; Ipson, S.; and Al-Fahdawi, S. 2018. A multimodal deep learning framework using local feature representations for face recognition. *Machine Vision and Applications*, 29(1): 35–54.

Andrews, M.; and Vigliocco, G. 2010. The hidden Markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, 2(1): 101–113.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Baltrušaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2): 423–443.

Barros, P.; Churamani, N.; Lakomkin, E.; Siqueira, H.; Sutherland, A.; and Wermter, S. 2018. The OMG-emotion behavior dataset. In *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Bishop, C. M.; and Nasrabadi, N. M. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3: 993–1022.

Cavallo, A.; Romeo, L.; Ansuini, C.; Battaglia, F.; Nobili, L.; Pontil, M.; Panzeri, S.; and Becchio, C. 2021. Identifying the signature of prospective motor control in children with autism. *Scientific reports*, 11(1): 1–8.

Chen, N.; Zhu, J.; Sun, F.; and Xing, E. P. 2012. Large-margin predictive latent subspace learning for multiview data analysis. *IEEE transactions on pattern analysis and machine intelligence*, 34(12): 2365–2378.

Correa, N. M.; Adali, T.; Li, Y.-O.; and Calhoun, V. D. 2010. Canonical correlation analysis for data fusion and group inferences. *IEEE signal processing magazine*, 27(4): 39–50.

Gao, J.; Li, P.; Chen, Z.; and Zhang, J. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5): 829–864.

Heo, J.; Lee, H. B.; Kim, S.; Lee, J.; Kim, K. J.; Yang, E.; and Hwang, S. J. 2018. Uncertainty-aware attention for reliable interpretation and prediction. *arXiv preprint arXiv:1805.09653*.

Hori, C.; Hori, T.; Lee, T.-Y.; Zhang, Z.; Harsham, B.; Hershey, J. R.; Marks, T. K.; and Sumi, K. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, 4193–4202.

Hsu, W.-N.; Zhang, Y.; and Glass, J. 2017. Unsupervised learning of disentangled and interpretable representations from sequential data. *arXiv preprint arXiv:1709.07902*.

Koirala, A.; Yu, Z.; Schiltz, H.; Van Hecke, A.; Armstrong, B.; and Zheng, Z. 2021. A Preliminary Exploration of Virtual Reality-Based Visual and Touch Sensory Processing Assessment for Adolescents With Autism Spectrum Disorder. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29: 619–628.

Kumar, K.; Sahu, S.; Majumdar, A.; and Chandra, M. G. 2021. AutoFuse: A Semi-supervised Autoencoder based Multi-Sensor Fusion Framework. In *2021 International Joint Conference on Neural Networks (IJCNN)*, 1–7. IEEE.

Lin, W.; Chu, H.; Wu, J.; Sheng, B.; and Chen, Z. 2013. A heat-map-based algorithm for recognizing group activities in videos. *TCSVT*, 23(11): 1980–1992.

Ma, G.; Yang, X.; Zhang, B.; and Shi, Z. 2016. Multi-feature fusion deep networks. *Neurocomputing*, 218: 164–171.

Mcauliffe, J. D.; and Blei, D. M. 2008. Supervised topic models. In *Advances in neural information processing systems*, 121–128.

Nakamura, T.; Araki, T.; Nagai, T.; and Iwahashi, N. 2011. Grounding of word meanings in latent dirichlet allocation-based multimodal concepts. *Advanced Robotics*, 25(17): 2189–2206.

Pantic, M.; Pentland, A.; Nijholt, A.; and Huang, T. S. 2007. Human computing and machine understanding of human behavior: A survey. In *Artifical intelligence for human computing*, 47–71. Springer.

Rasouli, A.; Kotseruba, I.; and Tsotsos, J. K. 2017. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 206–213.

Sano, A.; Chen, W.; Lopez-Martinez, D.; Taylor, S.; and Picard, R. W. 2018. Multimodal ambulatory sleep detection using LSTM recurrent neural networks. *IEEE journal of biomedical and health informatics*, 23(4): 1607–1617.

Seo, S.; Chan, H.; Brantingham, P. J.; Leap, J.; Vayanos, P.; Tambe, M.; and Liu, Y. 2018. Partially Generative Neural Networks for Gang Crime Classification with Partial Information. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 257–263. ACM.

Song, Y.; Morency, L.-P.; and Davis, R. 2012. Multimodal human behavior analysis: learning correlation and interaction across modalities. In *Proceedings of the 14th ACM international conference on Multimodal interaction*, 27–30.

Sorber, L.; Van Barel, M.; and De Lathauwer, L. 2015. Structured data fusion. *IEEE Journal of Selected Topics in Signal Processing*, 9(4): 586–600.

Srivastava, N.; and Salakhutdinov, R. R. 2012. Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, 2222–2230.

Tsai, Y.-H. H.; Liang, P. P.; Zadeh, A.; Morency, L.-P.; and Salakhutdinov, R. 2019. Learning Factorized Multimodal Representations. In *International Conference on Representation Learning*.

Tu, C.; Zhang, W.; Liu, Z.; Sun, M.; et al. 2016. Max-margin deepwalk: Discriminative learning of network representation. In *IJCAI*, volume 2016, 3889–3895.

Velivelli, A.; and Huang, T. S. 2008. Automatic video annotation using multimodal Dirichlet process mixture model. In *2008 IEEE International Conference on Networking, Sensing and Control*, 1366–1371.

Wang, Y.; and Mori, G. 2011. Max-margin Latent Dirichlet Allocation for Image Classification and Annotation. In *BMVC*, volume 2, 7.

Wu, M.; and Goodman, N. 2018. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, 5575–5585.

Zheng, E.; Yu, Q.; and Zheng, Z. 2023. Appendix: Sparse Maximum Margin Learning From Multimodal Human Behavioral Patterns. https://github.com/ritmininglab/SM2-MRS/. Accessed: 2023-02-01.

Zhou, T.; Thung, K.-H.; Zhu, X.; and Shen, D. 2019. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. *Human brain mapping*, 40(3): 1001–1016.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2012. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(Aug): 2237–2278.

Zhu, J.; and Xing, E. P. 2009. Maximum entropy discrimination Markov networks. *Journal of Machine Learning Research*, 10(Nov): 2531–2569.