Summer 8-6-2024

Bayesian Variational Inference in Keyword Identification and Multiple Instance Classification

Yaofang Hu
yaofangh@smu.edu

BAYESIAN VARIATIONAL INFERENCE IN KEYWORD IDENTIFICATION AND MULTIPLE INSTANCE CLASSIFICATION

Approved by:

Dr. Xinlei Wang
Professor in Department of Mathematics, UTA

Dr. Daniel F. Heitjan
Professor in Department of Statistics and Data Science, SMU & Peter O'Donnell Jr. School of Public Health, UTSW

Dr. Chul Moon
Assistant Professor in Department of Statistics & Data Science, SMU

Dr. Yichen Cheng
Associate Professor in Robinson College of Business, GSU

BAYESIAN VARIATIONAL INFERENCE IN KEYWORD IDENTIFICATION AND MULTIPLE INSTANCE CLASSIFICATION

A Dissertation Presented to the Graduate Faculty of the

Dedman College

Southern Methodist University

in

Partial Fulfillment of the Requirements

for the degree of

Doctor of Philosophy

with a

Major in Statistical Science

by

Yaofang Hu

B.S., Statistics, National University of Singapore

August 6, 2024

Acknowledgments

First of all, I would like to thank my advisor Dr. Xinlei Wang for her unwavering guidance, patience, support, and encouragement throughout my Ph.D. study. It has been an honor and privilege to learn from her. Her knowledge, vision, and passion for research have been a constant source of inspiration to me. Dr. Wang shows me the beauty of Bayesian statistics, shapes my research, and fosters my future career. It is her belief in me that encourages me to forward at every step. I could not ask for a better advisor.

I am deeply grateful to my collaborators, Dr. Yusen Xia and Dr. Yichen Cheng from Georgia State University. Their help and insightful suggestions are instrumental in finishing this thesis.

In addition, I would like to express my sincere gratitude to my committee members. I extend my thanks to Dr. Daniel Heitjan for his consistent support over the past four years and his constructive feedback on my thesis. I am also grateful to Dr. Moon for providing helpful feedback and suggestions for refining this work.

I would like to thank all of the great faculty members I met at SMU, for their excellent teaching and help. I thank my friends, who always stand by my side and give me mental support.

Last but not least, I would like to thank my family. Without their unconditional love, I would not be where I am today. They have sacrificed so much to help me pursue my dreams, and I am incredibly proud to be their child.

Hu, Yaofang B.S., Statistics, National University of Singapore

Bayesian Variational Inference in Keyword Identification and Multiple Instance Classification

Advisor: Dr. Xinlei Wang
Doctor of Philosophy degree conferred August 6, 2024
Dissertation completed May 28, 2024

This dissertation investigates (1) Variational Bayesian Semi-supervised Keyword Extraction and (2) Variational Bayesian Multimodal Multiple Instance Classification.

The expansion of textual data, stemming from various sources such as online product reviews and scholarly publications on scientific discoveries, has created a demand for the extraction of succinct yet comprehensive information. As a result, in recent years, efforts have been spent in developing novel methodologies for keyword extraction. Although many methods have been proposed to automatically extract keywords in the contexts of both unsupervised and fully supervised learning, how to effectively use partially observed keywords, such as author-specified keywords, remains an under-explored area. In Chapter 1, we propose a novel variational Bayesian semi-supervised (VBSS) keyword extraction approach, built on a recent Bayesian semi-supervised (BSS) technique that uses the information from a small set of known keywords to identify previously undetected ones. Our proposed VBSS method greatly enhances the computational efficiency of BSS via mean-field variational inference, coupled with data augmentation, which brings closed-form solutions at each step of the optimization process. Further, our numerical results show that VBSS offers enhanced accuracy for long texts and improved control over false discovery rates when compared with a list of state-of-the-art keyword extraction methods.

In Chapter 2, we apply mean-field variational inference on multiple instance learning (MIL). In MIL, objects are represented by bags of instances. Each instance shares the same feature set but has unique feature values. MIL aims to train models that predict bag-level outcomes based on these instances, making it a weakly supervised approach due to the lack of instance-level labels. While MIL methods focusing on binary classification are abundant, they often cannot identify which specific instances drive bag labels and have limited or little interpretability. Xiong et al. (2024) introduced MICProB, a Bayesian multiple instance classification (MIC) algorithm that addresses these issues. However, MICProB is computationally intensive and best suited for unimodal instances. To overcome these limitations, we propose a novel variational Bayesian multimodal MIC (vMMIC) algorithm. vMMIC handles diverse instance types and significantly improves computational efficiency through Bayesian variational inference, combined with data augmentation. We benchmark vMMIC against MICProB and many other MIC approaches on both simulated and real-world data. Results demonstrate vMMIC's superior performance, computational efficiency, and interpretability.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I dedicate this dissertation to my family.

Chapter 1

VARIATIONAL BAYESIAN SEMI-SUPERVISED KEYWORD EXTRACTION

1.1. Introduction

The growth of big data in recent years has resulted in an influx of information, leaving individuals susceptible to its volume. Consequently, there arises a need to distill the essence of this information, one aspect of which is to identify a collection of keywords to efficiently capture and succinctly summarize the core concepts conveyed within the text. This problem has attracted efforts from researchers, partially due to its practical values. For instance, extracting keywords from texts in platforms such as TripAdvisor and Airbnb is beneficial for enhancing recommendation accuracy; and capturing significant words from a news article enables readers to quickly decide whether to proceed with reading the article.

Various approaches have been developed for extracting keywords from textual data, which can be grouped into three main categories based on the availability of labeled texts: supervised, unsupervised, and semi-supervised methods [1]. Supervised methods use a collection of training articles with keywords labeled to train the algorithms. Examples of such methods include Hulth [2], Caragea et al. [3], and Bordoloi et al. [4]. These approaches can achieve high accuracy, thanks to the use of high-quality labeled data, which, in turn, requires considerable human effort to obtain. As a result, the practical applicability of supervised methods is limited by this data labeling requirement.

In contrast, unsupervised methods do not require any training data and can be further divided into two branches: graph-based or statistic-based. Graph-based techniques represent documents as graphs, with words being nodes, and edges representing some relationship

between nodes such as co-occurrence, syntax, or semantics [5]. TextRank (TR) [6] is the first graph-based algorithm. The key idea of TR is to transform a document into a graph and compute the importance score $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ based on the idea of PageRank [7], which is Google's famous algorithm on ranking webpages. Intuitively, TR assumes words which have frequent occurrences or co-occur with other important words are more likely to be important. Formally, TR finds the importance scores that solves the equation $\boldsymbol{\theta} = (1-d)\,\mathbf{1}_n + d\boldsymbol{G}^T\boldsymbol{\theta}$ for $\boldsymbol{\theta}$, where n is the number of candidate words, d is a damping parameter usually set to be 0.85, $\mathbf{1}_n$ is a vector of n 1's, $\boldsymbol{G} = \boldsymbol{D}^{-1}\boldsymbol{A}$ is the normalized adjacency matrix, $\boldsymbol{A}$ is the weighted adjacency matrix whose (i,j)-th entry represents the relation between the i-th and j-th words, and $\boldsymbol{D}$ is the degree matrix whose diagonals equal the row sums of $\boldsymbol{A}$ and off-diagonals are 0's. For instance, the (i,j)-th entry of $\boldsymbol{A}$ derived via co-occurrence rules is the number of co-occurrence of the i-th and j-th words within a fixed-width window. Following the success of TR, several subsequent approaches were introduced by varying how the graph is generated, such as TopicRank (TpR) [8] which represents topics as nodes and semantic relations as edges, PositionRank (PosR) [9] which incorporates the position of words into their importance scores and MultipartiteRank (MR) [10] which uses both topic information and position information. Statistic-based methods rank the importance of words via different statistics. Examples of such methods include TF-IDF [11] that equals the product of term frequency and inverse document frequency, KP-Miner [12] which combines TF-IDF and other factors such as position information, and YAKE [13] which relies on various factors such as word frequencies and position. Unsupervised methods offer greater applicability and flexibility in real-world situations as they do not rely on expensive labeled data. However, they are more susceptible to noise and can produce less desirable results when compared to supervised techniques.

To strike a balance between the cost of labeling texts and the accuracy of keyword identification, researchers have resorted to the third branch, the semi-supervised approaches, with the assumption that a small subset of keywords is known in advance. Consequently,

how to incorporate such partial information plays a key role for such methods. We point out that, in practice, this subset can be obtained from various sources, such as hashtags in Twitter posts or a limited number of keywords specified by authors in academic papers. Although the research on semi-supervised methods has gained attention in recent years, there is relatively limited work compared to supervised and unsupervised ones. Among them, Li et al. [14] developed an interesting method (labeled SS) that integrates the partial label information into the calculation of importance scores while preserving the so-called local consistency [15] by solving $\boldsymbol{\theta} = (1-d)\,\boldsymbol{y} + d\boldsymbol{Q}^T\boldsymbol{\theta}$ for $\boldsymbol{\theta}$ (equivalently, by minimizing $\sum_{i,j} A_{ij}\left(\frac{1}{\sqrt{D_{ii}}}\theta_i - \frac{1}{\sqrt{D_{jj}}}\theta_j\right)^2 + (1-d)/d \parallel \boldsymbol{\theta} - \boldsymbol{y} \parallel^2$). Here, $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is a vector of observed labels, $y_i = 1$ if the i-th word is observed to be a keyword and 0 otherwise; and $\boldsymbol{Q} = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$ is another version of the normalized adjacency matrix, with $\boldsymbol{D}$ and $\boldsymbol{A}$ defined in the previous paragraph. A penalty is placed on the distance between observed label $\boldsymbol{y}$ and the importance scores $\boldsymbol{\theta}$ to make sure other words learn their importance scores from the observed words. As another state-of-the-art semi-supervised keyword extraction approach, BSS [1] integrated the importance scores $\boldsymbol{\theta}$ proposed by TR into a Bayesian logistic regression model and uses the partial label information $\boldsymbol{y}$ to formulate the likelihood function. Ye and Wang [16] and Jonathan and Karnalim [17] also fall into the category of semi-supervised methods, but they differ significantly from SS and BSS approaches and assume that within a collection of documents, a small proportion of the documents are fully labeled while the remainder are unlabeled. In this paper, like SS and BSS, we assume for each document, a subset of the keywords is known. Consequently, we exclude these two methods from our comparative study. We also refer readers to Wang et al. [1] for more in-depth information on each selected benchmark method.

Semi-supervised methods offer advantages to both supervised and unsupervised ones. However, the aforementioned approaches have their own limitations. For example, SS requires manual threshold determination for selecting words with top importance scores. On the other hand, while BSS is fully automatic in that manner, it is computationally inten-

sive and less suitable for long articles. To address these limitations, we propose a novel algorithm called Variational Bayesian Semi-supervised Keyword Extraction (VBSS). Built on BSS, VBSS introduces key distinctions in both computation and model setup. In terms of computation, VBSS leverages mean-field variational inference (VI) [18, 19] for posterior approximation and adopts the technique of data augmentation [20] to guarantee closed forms for all parameters, which significantly improve the computation efficiency. In contrast, BSS relies on a computationally intensive MCMC sampling procedure. In terms of model setup, VBSS incorporates the probit function to link the importance scores $\boldsymbol{\theta}$, and the probabilities of being a keyword, denoted by $\boldsymbol{p} = (p_1, \cdots, p_n)^T$. Furthermore, we introduce additional parameters to enhance the flexibility of our model for improved accuracy. As a result, VBSS demonstrates notable performance improvements over BSS in our data examples. It exhibits enhanced control over the false discovery rate (FDR) and computational efficiency, for both short and long articles. In addition, VBSS excels in efficiently handling long full-text articles with superior performance in keyword identification.

The remainder of this chapter is organized as follows. Section 1.2 introduces the construction of the Bayesian model. In Section 1.3, we provide a detailed description of the VBSS algorithm, including computation and the pseudo code. We compare our VBSS and existing methods using real-world data sets in Section 1.4. Section 1.5 summarizes our work and discusses future directions.

1.2. Bayesian Hierarchical Modeling

In our semi-supervised model setting, we assume that only a subset of the actual keywords in an article is known (or observed), meaning each candidate word (say, the i-th) is associated with two labels: the observed label y_i and the true label y_i^*, $i = 1, \ldots, n$. The observed label $y_i = 1$ indicates the i-th term is known (or observed) to be a keyword and 0 otherwise, and the (unknown) true label $y_i^* = 1$ if the i-th term is indeed a keyword and 0 otherwise. Here, only true keywords can be observed, meaning that the words observed to be keywords must

be true ones (i.e., $y_i^* = 1$ indicates $y_i = 1$). On the other hand, there is a (large) possibility that true keywords are not observed; that is, given $y_i = 0$, y_i^* can take values of 0 or 1. Let α_i denote the conditional probability that the i-th candidate word is not observed to be a keyword given it actually is, and $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^T$ denote the vector of these conditional probabilities. Thus,

$$\begin{cases} P(y_i = 0 | y_i^* = 1) = \alpha_i, \\ P(y_i = 0 | y_i^* = 0) = 1. \end{cases} \tag{1.1}$$

For word i, to link its probability of being a keyword to its importance score, we assume $p_i = P\left(y_i^* = 1\right) = \Phi\left(a + b\theta_i\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of $N(0, 1)$, and a and b represent the intercept and slope parameters in this generalized linear model. The likelihood function is

$$p\left(\boldsymbol{y} \mid a, b, \boldsymbol{\theta}, \boldsymbol{\alpha}\right) = \prod_{i=1}^{n} \left[(1 - \alpha_i) \cdot \Phi\left(a + b\theta_i\right)\right]^{y_i} \left[1 - (1 - \alpha_i)\, \Phi\left(a + b\theta_i\right)\right]^{1 - y_i}.$$

It is worth noting that BSS assumes $p_i = \text{logit}^{-1}\left(\theta_i\right)$ instead. With two added parameters a and b, our proposed model gains increased flexibility. Furthermore, when switching from the logit link to the probit link, it facilitates efficient algorithm design by providing a closed-form solution at each step, as will be shown in Section 1.3.3. We point out that $\boldsymbol{\theta}$ represents the importance scores derived via different methods. Since methods can represent documents in graphs and calculate the importance scores in different ways, the scales of different $\boldsymbol{\theta}$ might vary. However, they carry the same conceptual meaning, which is the relative importance of candidate words.

To incorporate the graph structure of the article and the observed label information, we consider a multivariate normal prior on $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$:

$$\pi\left(\boldsymbol{\theta} | \sigma^2\right) = N\left(\boldsymbol{\theta_0}, \boldsymbol{B}^{-1}\left(\boldsymbol{B}^{-1}\right)^T \sigma^2\right), \tag{1.2}$$

where $\boldsymbol{\theta_0} = (1-d)\,\boldsymbol{B}^{-1}\boldsymbol{y}$ (i.e., the solution from Li et al. [14]) and $\boldsymbol{B} = \boldsymbol{I} - d\boldsymbol{Q}^T$, with $\boldsymbol{Q}$ and d defined in the introduction and $\boldsymbol{I}$ being the $n \times n$ identity matrix. For simplicity, we denote $\boldsymbol{B}^{-1}(\boldsymbol{B}^{-1})^T$ by $\boldsymbol{U}$. We use θ_{0i} to denote the i-th element of $\boldsymbol{\theta}_0$, and $\boldsymbol{U}_{ii}$ to denote the i-th diagonal element of $\boldsymbol{U}$.

We set the priors of a and b to be $N\left(0, \sigma_a^2\right)$ and $N\left(0, \sigma_b^2\right)$, respectively. We choose an inverse gamma prior $IG\left(\tau, \tau\right)$ for σ^2, and a uniform prior $\pi\left(\alpha_i\right) = 1$ for α_i, $\alpha_i \in [0, 1]$. Note that σ_a^2, σ_b^2 and τ are user-defined hyperparameters. We suggest using diffuse or vague priors so we set $\sigma_a^2 = \sigma_b^2 = 10$ and $\tau \in (0.01, 0.1, 1)$.

1.3. A Variational Bayesian Approach

1.3.1. Preliminaries for variational inference

The gist of VI is to find a joint probability density function (within a candidate approximation density family) that best approximates the posterior distributions of parameters of interest in terms of Kullback–Leibler (KL) divergence. Without loss of generality, for this subsection only, let's suppose we have parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$ and data $\boldsymbol{y} = (y_1, \ldots, y_n)^T$. We use $\mathscr{Q}$ to denote the family of the candidate approximation densities $(q\left(\boldsymbol{\theta}\right))$ to the exact posterior distribution of $\boldsymbol{\theta}$ $(p\left(\boldsymbol{\theta}|\boldsymbol{y}\right))$. The "best" candidate $q^*\left(\boldsymbol{\theta}\right)$ is the one that minimizes its $\mathcal{KL}$-divergence to the posterior $p(\boldsymbol{\theta}|\boldsymbol{y})$. That is, $q^*(\boldsymbol{\theta}) =_{q(\boldsymbol{\theta})\in\mathscr{Q}} \mathcal{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y}))$. It turns out that the $\mathcal{KL}$-divergence is intractable since

$$
\begin{aligned}
\mathcal{KL}\left(q\left(\boldsymbol{\theta}\right)||p\left(\boldsymbol{\theta}|\boldsymbol{y}\right)\right) &= -\int q(\boldsymbol{\theta}) \ln\left(\frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q\left(\boldsymbol{\theta}\right)}\right) d\boldsymbol{\theta} \\
&= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln q\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln p\left(\boldsymbol{\theta} \mid \boldsymbol{y}\right)\right] \\
&= \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln q\left(\boldsymbol{\theta}\right)\right] - \mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln p\left(\boldsymbol{\theta}, \boldsymbol{y}\right)\right] + \ln p\left(\boldsymbol{y}\right),
\end{aligned}
$$

which involves the term $\ln p\left(\boldsymbol{y}\right)$ [18]. Here, $\mathbb{E}_{q(\boldsymbol{\theta})}(\cdot)$ denotes taking the expectation with respect to $q(\boldsymbol{\theta})$. Since $\ln p\left(\boldsymbol{y}\right)$ does not depend on q and can be treated as a constant,

we can maximize an alternative quantity $\mathbb{E}_{q(\boldsymbol{\theta})}\left[\ln p\left(\boldsymbol{\theta}, \boldsymbol{y}\right) - \ln q\left(\boldsymbol{\theta}\right)\right]$, known as the evidence lower bound (ELBO), which equals to a constant minus $\mathcal{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y}))$. We follow the mean-field VI machinery, and assume the variational distribution over $\boldsymbol{\theta}$ can be factorized as $q(\theta_1, \ldots, \theta_n) = \prod_{i=1}^{n} q(\theta_i)$. We note that the mean-field VI is closely related to the approximation framework of the mean field theory [21] in physics. The coordinate ascent algorithm is often used to maximize the ELBO by iteratively updating the variational distribution $q\left(\theta_i\right)$ while holding $q\left(\theta_{-i}\right)$ (i.e., all other variational distributions) fixed. In each iteration, it can be shown that $q^*\left(\theta_i\right) \propto \exp\left(\mathbb{E}_{-\theta_i}\left[\log\left(p\left(\theta_i \mid \theta_{-i}, y\right)\right)\right]\right)$, which is obviously reminiscent of Gibbs sampling [22].

1.3.2. Data augmentation

With the current model setup, not every variational distribution has closed-form updates in the mean-field VI stage, which makes the computation difficult. To solve this problem, we follow the idea of Albert and Chib [20] and introduce latent variables $\boldsymbol{z} = (z_{1,} \ldots z_n)^T$, whose signs determine the value of $\boldsymbol{y}^*$; that is, the true label y_i^* is 1 if $z_i > 0$ and 0 if $z_i \leq 0$, and $z_i \sim N(a + b\theta_i, 1)$. Then we have

$$
\begin{aligned}
P(y_i|z_i, \alpha_i) &= P(y_i, y_i^* = 1|z_i, \alpha_i) + P(y_i, y_i^* = 0|z_i, \alpha_i) \\
&= \begin{cases} (1 - \alpha_i) \cdot P(y_i^* = 1|z_i), & \text{if } y_i = 1 \\ \alpha_i \cdot P(y_i^* = 1|z_i) + P(y_i^* = 0|z_i), & \text{if } y_i = 0 \end{cases} \\
&= \left[(1 - \alpha_i) \cdot 1(z_i > 0)\right]^{y_i} \left[\alpha_i \cdot 1(z_i > 0) + 1(z_i \leq 0)\right]^{1-y_i}.
\end{aligned}
$$

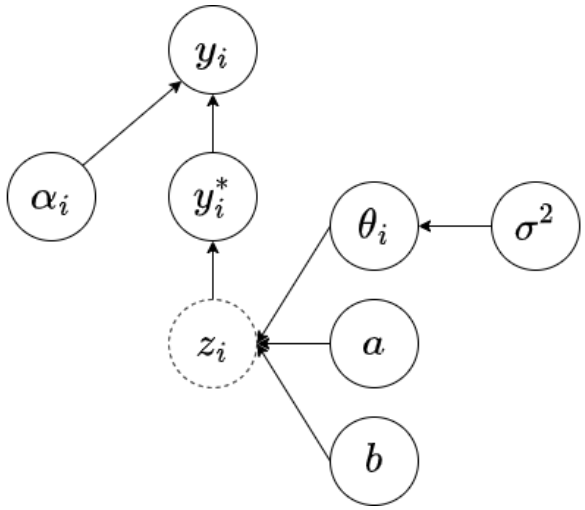


Figure 1.1: The hierarchical structure of the proposed VBSS model.

A diagram is shown in Figure 1.1 to illustrate the hierarchical structure of our proposed VBSS model, where the auxiliary variable z_i is shown in a dashed circle to indicate it is introduced to facilitate computation.

1.3.3. Variational Bayesian semi-supervised keyword extraction

With the introduction of the latent variables $\boldsymbol{z} = (z_1, \ldots z_n)^T$, the joint posterior distribution becomes

$$
\begin{aligned}
p\left(\boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha} | \boldsymbol{y}\right) &\propto p\left(\boldsymbol{y} | \boldsymbol{z}, \boldsymbol{\alpha}\right) p\left(\boldsymbol{z} | \boldsymbol{\theta}, a, b\right) \pi\left(\boldsymbol{\theta} | \sigma^2\right) \pi\left(a\right) \pi\left(b\right) \pi\left(\sigma^2\right) \pi\left(\boldsymbol{\alpha}\right) \\
&= \left\{ \prod_{i=1}^{n} p\left(y_i | z_i, \alpha_i\right) p\left(z_i | \theta_i, a, b\right) \pi\left(\alpha_i\right) \right\} \pi\left(\boldsymbol{\theta} | \sigma^2\right) \pi\left(a\right) \pi\left(b\right) \pi\left(\sigma^2\right) \\
&= \prod_{i=1}^{n} \left\{ \left[(1-\alpha_i) \cdot 1\left(z_i > 0\right) \right]^{y_i} \left[\alpha_i \cdot 1\left(z_i > 0\right) + 1\left(z_i \leq 0\right) \right]^{1-y_i} \cdot N\left(a + b\theta_i, 1\right) \right\} \\
&\quad \cdot N\left(\boldsymbol{\theta_0}, \boldsymbol{U}\sigma^2\right) \cdot N\left(0, \sigma_a^2\right) \cdot N\left(0, \sigma_b^2\right) \cdot IG\left(\tau, \tau\right).
\end{aligned}
$$

Now, our task is to find the variational distributions such that

$$
p\left(\boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha} | \boldsymbol{y}\right) \approx \left(\prod_{i=1}^{n} q(z_i) \right) \left(\prod_{i=1}^{n} q(\theta_i) \right) q(a) q(b) q(\sigma^2) \left(\prod_{i=1}^{n} q(\alpha_i) \right).
$$

The functional forms can be found through the optimization of the variational distributions [18]. By taking the derivative with respect to $q(\theta_i)$ and setting the derivative to zero, we have:

$$\begin{aligned}
\ln\, q^*\left(\theta_i\right) &= \mathbb{E}_{q^*(-\theta_i)}\left\{\ln\left[p\left(\boldsymbol{y}|\boldsymbol{z},\boldsymbol{\alpha}\right)p\left(\boldsymbol{z}|\boldsymbol{\theta},a,b\right)\pi\left(\boldsymbol{\theta}|\sigma^2\right)\pi\left(a\right)\pi\left(b\right)\pi(\sigma^2)\pi\left(\boldsymbol{\alpha}\right)\right]\right\}+\text{const} \\
&= \mathbb{E}_{q^*(z_i,a,b)}\left[\ln p\left(z_i|\theta_i,a,b\right)\right]+\mathbb{E}_{q^*(\sigma^2)}\left[\ln\pi\left(\theta_i|\sigma^2\right)\right]+\text{const} \\
&= \mathbb{E}_{q^*(z_i,a,b)}\left[-\frac{1}{2}\left(z_i-(a+b\theta_i)\right)^2\right]+\mathbb{E}_{q^*(\sigma^2)}\left[-\frac{1}{2\boldsymbol{U}_{ii}\sigma^2}\left(\theta_i-\theta_{0i}\right)^2\right]+\text{const} \\
&= -\frac{1}{2}\left(\mathbb{E}_{q^*(b)}\left[b^2\right]+\frac{1}{\boldsymbol{U}_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\right)\theta_i^2 \\
&\quad+\left(\mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(z_i)}\left[z_i\right]-\mathbb{E}_{q^*(a)}\left[a\right]\mathbb{E}_{q^*(b)}\left[b\right]+\frac{1}{\boldsymbol{U}_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\theta_{0i}\right)\cdot\theta_i+\text{const},
\end{aligned}$$

which is the log of the probability density function of a normal distribution. By completing the square term, we obtain

$$\begin{aligned}
q^*(\theta_i) &= N(\mu_{\theta_i},S_{\theta_i}), \\
\mu_{\theta_i} &= S_{\theta_i}\cdot\left(\mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(z_i)}\left[z_i\right]-\mathbb{E}_{q^*(a)}\left[a\right]\mathbb{E}_{q^*(b)}\left[b\right]+\frac{1}{\boldsymbol{U}_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\theta_{0i}\right), \\
S_{\theta_i} &= \left(\mathbb{E}_{q^*(b)}\left[b^2\right]+\frac{1}{\boldsymbol{U}_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\right)^{-1}.
\end{aligned}$$

Similarly, we derive the variational distributions of z_i, a, b, σ^2, and α_i, respectively. The details are provided in Section A.1 in Appendix A. As shown in the derivation for θ_i above, the optimized variational distribution of a certain parameter involves the expectations of other parameters listed below, which can be obtained straightforwardly based on each parameter's variational distribution:

- $\mathbb{E}_{q^*(\sigma^2)}\left[1/\sigma^2\right]=(\tau+\frac{n}{2})/(\tau+g)$, where $g=\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)^T\frac{\boldsymbol{U}^{-1}}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta}_0\right)\right]$.
- $\mathbb{E}_{q^*(\theta_i)}[\theta_i]=\mu_{\theta_i}=S_{\theta_i}\cdot\left(\mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(z_i)}\left[z_i\right]-\mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(a)}\left[a\right]+\frac{1}{\boldsymbol{U}_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\theta_{0i}\right)$, where $S_{\theta_i}=\left(\mathbb{E}_{q^*(b)}\left[b^2\right]+\frac{1}{U_{ii}}\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right]\right)^{-1}$.

- $\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}^T\boldsymbol{\theta}\right] = \boldsymbol{\mu_\theta}^T\boldsymbol{\mu_\theta} + tr(\boldsymbol{S_\theta})$.
- $g = \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[(\boldsymbol{\theta}-\boldsymbol{\theta_0})^T\frac{\boldsymbol{U}^{-1}}{2}(\boldsymbol{\theta}-\boldsymbol{\theta_0})\right] = tr\left(\frac{\boldsymbol{U}^{-1}}{2}\cdot\boldsymbol{S_\theta}\right) + \left(\mathbb{E}_{q^*(\boldsymbol{\theta})}[\boldsymbol{\theta}]-\boldsymbol{\theta_0}\right)^T\frac{\boldsymbol{U}^{-1}}{2}\left(\mathbb{E}_{q^*(\boldsymbol{\theta})}[\boldsymbol{\theta}]-\boldsymbol{\theta_0}\right)$.
- $\mathbb{E}_{q^*(a)}[a] = \frac{\sigma_a^2}{n\sigma_a^2+1}\cdot\left(\mathbb{E}_{q^*(\boldsymbol{z})}[\boldsymbol{z}]^T\mathbf{1}_n - \mathbb{E}_{q^*(b)}[b]\,\mathbb{E}_{q^*(\boldsymbol{\theta})}[\boldsymbol{\theta}]^T\mathbf{1}_n\right)$.
- $\mathbb{E}_{q^*(b)}[b] = \left(\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}^T\boldsymbol{\theta}\right] + \frac{1}{\sigma_b^2}\right)^{-1}\cdot\left(\mathbb{E}_{q^*(\boldsymbol{\theta})}[\boldsymbol{\theta}]^T\mathbb{E}_{q^*(\boldsymbol{z})}[\boldsymbol{z}] - \mathbb{E}_{q^*(b)}[a]\cdot\mathbb{E}_{q^*(\theta)}[\boldsymbol{\theta}]^T\mathbf{1}_n\right)$.
- $\mathbb{E}_{q^*(\alpha_i)}[\alpha_i] = \dfrac{\mathbb{E}_{q^*(z_i)}[(1-y_i)\,1(z_i>0)]+1}{y_i+\mathbb{E}_{q^*(z_i)}[(1-y_i)\,1(z_i>0)]+2}$.
- $\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i] = \psi\left(\mathbb{E}_{q^*(z_i)}[(1-y_i)\,1(z_i>0)]+1\right) - \psi\left(y_i+\mathbb{E}_{q^*(z_i)}[(1-y_i)\,1(z_i>0)]+2\right)$, where ψ is the digamma function.
- If $y_i = 1$, $\mathbb{E}_{q^*(z_i)}[z_i] = m_i + \dfrac{\phi(m_i)}{1-\Phi(-m_i)}$, where $\phi(\cdot)$ is the standard normal density; if $y_i = 0$,

$$\mathbb{E}_{q^*(z_i)}[z_i] = \frac{m_i\cdot t_i - \phi(m_i) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}\cdot[m_i\cdot(1-t_i)+\phi(m_i)]}{t_i+(1-t_i)e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}.$$

  Here, $m_i = \mathbb{E}_{q^*(a)}[a] + \mathbb{E}_{q^*(b)}[b]\cdot\mathbb{E}_{q^*(\theta_i)}[\theta_i]$ and $t_i = \Phi(-m_i)$.

With the expectations shown above, we can update the variational posterior distributions iteratively. At convergence, the final output is a vector of estimated probability of each word being a keyword $\hat{\boldsymbol{p}} = (\hat{p}_1,\ldots,\hat{p}_n)^T$, where $\hat{p}_i = \Phi(\hat{m}_i)$, where $\hat{m}_i$ is evaluated using last updated values of a, b and θ_i.

To make the final decision of whether a candidate word is a keyword or not, following BSS, we adopt an FDR control machinery [23] to set the threshold, and words with probabilities larger than that will be selected as keywords. Given a probability threshold h, the estimated FDR is

$$\widehat{FDR}(h) = \frac{\sum_{i=1}^n((1-\hat{p}_i)1(\hat{p}_i\geq h))}{\sum_{i=1}^n 1(\hat{p}_i\geq h)}.$$

With a pre-specified FDR cutoff γ such as 0.05, 0.1, or 0.15, we select the largest h such that $\widehat{FDR}(h)\leq\gamma$.

Based on the output $\hat{\boldsymbol{p}} = (\hat{p}_1, \cdots, \hat{p}_n)^T$ of VBSS, we proceed to calculate the estimated FDR with h being each possible $\hat{p}_i$. Then, for a pre-specified γ, we aim to identify the smallest $\hat{p}_i$ (say, $\hat{p}_k$) such that $\widehat{FDR}(h) \leq \gamma$ is satisfied. Consequently, any candidate word with a probability larger than $\hat{p}_k$ is considered as a keyword. We emphasize that this FDR-based identification approach is automated, offering the flexibility to select a varying number or proportion of keywords from different articles. This sets both BSS and VBSS apart from many existing methods that rely on a fixed threshold, often selecting one-third of candidate words in an article as keywords. These conventional methods inherently make the unrealistic assumption of a uniform proportion of keywords across all documents.

1.3.4. Algorithm

Building upon the modeling and derivations presented in the preceding subsections, the pseudo code for VBSS is provided in Algorithm 1. The VBSS algorithm requires specifying the hyperparameter values for the priors including σ_a^2, σ_b^2 and τ, a convergence threshold, and the maximum number of iterations allowed.

To initiate the iterative computation of expectations of $\boldsymbol{z}$, $\boldsymbol{\theta}$, a, b, $\boldsymbol{\alpha}$, and σ^2, appropriate initial values of these expectations need to be determined. We use $\boldsymbol{\theta}_{0i}$ (the solution from SS) as the initial point of $\mathbb{E}_{q^*(\theta_i)}[\theta_i]$; and 0, 1 and 1 as the initial values of $\mathbb{E}_{q^*(a)}[a]$, $\mathbb{E}_{q^*(b)}[b]$ and $\mathbb{E}_{q^*(\sigma^2)}[1/\sigma^2]$, respectively; and to initialize $\mathbb{E}_{q^*(\alpha_i)}[\alpha_i]$, we suggest using a roughly estimated proportion for unobserved keywords in the article, typically between 0.5 and 0.7. Then VBSS will iteratively update the expectations based on what we obtained in Section 1.3.3, and check the convergence at each iteration. Based on our model, in the j-th iteration, the probability of being a keyword for the i-th candidate is estimated to be $\hat{p}_i^j = \Phi\left(\hat{m}_i^j\right)$, $j \in (1, \ldots, miter)$ and $i \in (1, \ldots, n)$. The algorithm stops when the probability vectors of two successive iterations are sufficiently close. The final output of VBSS is a vector of

probabilities: $\hat{\boldsymbol{p}} = (\hat{p}_1, \ldots, \hat{p}_n)^T$, the i-th element of which represents the possibility of the i-th candidate being a keyword.

1.4. Applications

We compare our proposed VBSS with existing methods in terms of performance on keyword identification and time consumption on real-world data examples with ground truths. We exclude supervised approaches from our comparison since they require a large amount of fully labeled texts for algorithm training. Additionally, unsupervised or semi-supervised methods that rely on other additional information, such as topics generated from topic modeling, are not considered. The methods included for comparison are three semi-supervised approaches (SS, BSS and VBSS) and seven unsupervised methods (TR, TpR, PosR, MR, TF-IDF, KPMiner, and YAKE), among which VBSS and BSS are Bayesian methods. For the implementation of the BSS method, we utilize the R [24] code provided in Wang et al. [1]. TpR, PosR, MR, TF-IDF, KPMiner, and YAKE are implemented using the python toolkit pke (python keyphrase extraction) developed by Boudin [25]. We develop our own R code to implement TR, SS, and VBSS, which is publicly available at github.com/YaofangHuYaofang/VBSS.

To evaluate the performance of the different methods, we measure their precisions, recalls, and F-measures for each dataset, defined as Precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$, recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$, and F-measure $= 2\times\mathrm{precision}\times\mathrm{recall}/(\mathrm{precision}+\mathrm{recall})$, where TP stands for true positives, FP for false positives, TN for true negatives, and FN for false negatives.

As mentioned before, VBSS utilizes an automatic FDR control procedure to select keywords, with the number of selected keywords determined by the pre-specified threshold γ, similar to BSS. In contrast, all the other methods output importance scores and require manual selection of a cutoff to determine the total number of identified keywords in each document. To ensure a fair comparison, we control the total number of keywords identified

Algorithm 1	Variational Bayesian Semi-supervised Keyword Extraction (VBSS)
Input	A document and a list of observed keywords; hyperparameters τ, σ_a^2 and σ_b^2; the damping factor d; the convergence threshold ϵ; the maximum number of iterations $miter$.
Step 1	Construct an undirected weighted graph from the input document.
Step 2	Specify initial values for expectations of $\boldsymbol{z}$, $\boldsymbol{\theta}$, a, b, σ^2 and $\boldsymbol{\alpha}$ at $j = 0$.

Step 3	**For** $j \in (1, \ldots, miter)$:
	update the expectations of $\boldsymbol{z}$, $\boldsymbol{\theta}$, a, b, σ^2 and $\boldsymbol{\alpha}$,
	based on expectations shown in Section 1.3.3.
	compute $\hat{p}_i^{(j)} = \Phi\left(\hat{m}_i^{(j)}\right)$ for $i \in (1, \ldots, n)$.
	If the average probability over candidate words between the j-th and $(j-1)$-th iteration $\sum_{i=1}^{n} \lvert\hat{p}_i^{(j)} - \hat{p}_i{}^{(j-1)}\rvert / n < \epsilon$,
	break
	else
	continue
	end if
	End for

Output	A vector of probabilities $\hat{\boldsymbol{p}} = (\hat{p_1}, \ldots, \hat{p}_n)^T$, the i-th element of which represents the possibility of the i-th candidate being a keyword.

for those methods that require manual selection to be the same as the total number of keywords identified by VBSS. To achieve this, given a corpus consisting of multiple documents, we calculate the proportion of keywords identified by VBSS across all documents among all candidates (say, $r\%$). Thereafter, for other methods which require manual selection, we select the $r\%$ (rounded) top-ranked words for each article.

The characteristics of three real-world benchmark datasets used for comparison, including the number of keywords, the number of candidate words, and the proportion of keywords, are detailed in Table 1.1. The PubMed dataset comprises long full-text articles in the field of biomedical research. The SemEval2010 dataset includes long full-text articles related to computing and information technology. In contrast, the Hulth dataset contains shorter abstracts, typically consisting of a single paragraph. These datasets provide a diverse range of text lengths and text topics, allowing us to assess the performance of the methods across different document types and documents from different fields.

1.4.1. PubMed articles

We first evaluate the performance of VBSS on long articles using the PubMed articles collected from the PubMed Central [26]. Each article in the dataset is equipped with pre-assigned keywords, which serve as the ground-truth for performance measures. Prior to our analyses, the articles and keywords undergo standard natural language processing (NLP) steps, including tokenization and part-of-speech tagging (POS-tagging). Using tokenization, we split sentences or paragraphs into individual words, which become the basic units for further analysis. POS-tagging assigns word units to different classes such as nouns, verbs, adverbs, etc., based on their corresponding part-of-speech. For the keyword extraction task, we remove stop words such as prepositions and conjunctions that provide limited information. Next, we apply a stemming process to convert words into their respective stem forms. For example, words such as confidence and confident are stemmed as confid. In addition, we

Table 1.1: The description and summary statistics of the three benchmark datasets. All three datasets have been preprocessed to remove stop words (i.e. words with little meanings). We converted all words into their stems and removed words appearing less than three times for datasets with long articles, such as PubMed and SemEval2010. We only keep texts with at least 11 keywords for all three datasets.

Dataset	Description		Min	Q_1	Q_2	Mean	Q_3	Max
PubMed	30 randomly	#keywords	11	11	14.5	15.2	17.8	25
[26]	selected from a library	#candidates	73	179	220.5	202.8	242.5	262
	of 500 biomedical papers	%keywords	4.9	6.3	6.9	8.0	8.3	17.8
SemEval2010	239 full papers	#keywords	11	17	20	20.82	24	39
[27]	from ACM	#candidates	128	272.5	305	308.1	334.5	550
	digital library	%keywords	2.7	5.5	6.8	6.9	8.1	15.2
Hulth	1,459 abstracts	#keywords	11	14	19	20.4	25	56
[2]	of computer	#candidates	14	45	57	58.5	71	151
	science papers	%keywords	11.3	27.8	36.2	36.7	44.5	78.6

eliminate words that appear only once or twice throughout the articles as they are unlikely to be keywords in a long article. We further randomly select a subset of 30 full-text papers with at least 11 keywords from the dataset, which, along with the above preprocessing procedures, allows us to apply the computationally intensive BSS method as well. In the processed documents, each word is considered a vertex in a graph, with edges representing co-occurrences between two candidates within a two-word window. The weights assigned to the edges correspond to the total number of co-occurrences observed throughout the article. For each of the 30 articles, we randomly select $k = 5$ keywords from the corresponding ground truth set and designate them as the observed keywords. For VBSS, the hyperparameters are set as $\tau = 0.1$ and $\sigma_a^2 = \sigma_b^2 = 10$. We further set the convergence threshold $\epsilon = 1 \times 10^{-10}$ and the maximum number of iterations $miter = 500$. For simplicity, we set $\alpha_i = \alpha$ and put a uniform prior on α (for detail, see Section A.2 in Appendix A). Default values are used for the hyperparameters of the other algorithms. Note that we use these settings across all subsequent numerical experiments. As mentioned earlier, for fair comparison, we control the total number of identified keywords across various approaches (except for BSS) to be

(approximately) the same. The total numbers of positives identified by VBSS and BSS with different γ (ranging from 0.05 to 0.3) are summarized in Table 1.2, where we find that BSS consistently identifies more positives than VBSS.

Semi-supervised techniques inherently leverage the partially available label information. Even when $\gamma = 0.05$, our VBSS successfully identifies all the observed keywords, and both BSS and SS effectively capture nearly all of the observed keywords. This holds true, not to mention when $\gamma > 0.05$. In contrast, unsupervised methodologies miss and fail to encompass this knowledge. Table 1.3 shows that among the 150 observed keywords, unsupervised approaches recover only a small fraction of them.

The performance comparison is shown in Table 1.4, with the highest precisions, recalls, and overall F-measures under different γ highlighted, and the lowest in italic. VBSS demonstrates the highest precision for almost every value of γ, indicating its tendency to identify fewer false positives compared to other approaches. On the other hand, BSS achieves the highest recall among all methods, suggesting that the labels generated by BSS have wider coverage. When considering overall performance, VBSS outperforms other approaches for most values of γ, followed by SS. Notably, BSS achieves the highest F-measure when γ is small (0.05) but falls behind when γ gets larger. Unsupervised methods exhibit poor performance across the board, evidently due to no use of of the label information. Further, in the last line of the table, we have included the total computational time (on a Windows 10 Operating System equipped with an Intel(R) Core(TM) i5-7400 CPU operating at 3.00GHz and 24.0 GB of RAM) for each method, we observe that VBSS runs significantly faster

Table 1.2: PubMed data: the total numbers of positives identified by VBSS and BSS with different γ.

FDR control γ	0.05	0.1	0.15	0.2	0.25	0.3
Total No. of Positives (VBSS)	150	181	210	244	305	388
Total No. of Positives (BSS)	211	249	295	350	431	555

Table 1.3: PubMed data: the total number of observed keywords that are successfully identified by various approaches.

FDR Control γ	VBSS	BSS	SS	TR	TpR	MR	PosR	TF-IDF	KPMiner	YAKE
0.05	150	100	135	19	12	9	16	20	17	15
0.1	150	127	140	21	15	13	16	23	21	17
0.15	150	138	146	23	16	16	19	26	23	18
0.2	150	147	146	25	17	16	20	28	26	20
0.25	150	150	148	30	23	18	21	31	30	23
0.3	150	150	150	38	27	20	22	37	37	33

than BSS, with its computational time comparable to many of the non-Bayesian benchmark methods.

We can observe from Table 1.3 that unsupervised approaches cannot ensure the identification of observed keywords, simply because they cannot utilize the labeled information. Thus, an alternative way of comparison is to force the observed keywords to be positive for those approaches so that they are not disadvantaged in the comparison. More specifically, for a document of n candidate words, 5 observed keywords are automatically set to be true keywords, among other $n-5$ words, top $n \times r\% - 5$ (rounded) candidate words with highest importance scores are identified as keywords. After such adjustment, the precisions, recalls, and F-measures of unsupervised methods TR, TpR, MR, PosR, TF-IDF, KPMiner, and YAKE show significant improvement when the observed keywords are forced to be positive. Furthermore, we observe the same general pattern as in Table 1.4. The detailed results are reported in Table 1.1 in Appendix A.

1.4.2. SemEval2010 data

SemEval2010 [27] is another benchmark dataset widely used in NLP-related studies, which consists of 244 full-text computer science papers. Each paper is associated with an author-assigned set of keywords and a professional-editors-assigned set of keywords. We

Table 1.4: PubMed data: the comparison of precisions, recalls, F-measures and time consumption measured of various keyword extraction approaches with different FDR control γ.

FDR cutoff γ						**Precision**				
	VBSS	BSS	SS	TR	TpR	MR	PosR	TF-IDF	KPMiner	YAKE
0.05	**1**	0.536	0.975	0.435	0.286	*0.252*	0.401	0.456	0.422	0.340
0.1	**0.934**	0.386	0.869	0.426	0.284	*0.230*	0.372	0.459	0.432	0.317
0.15	**0.848**	0.282	0.817	0.427	0.272	*0.221*	0.338	0.455	0.446	0.300
0.2	**0.783**	*0.205*	0.753	0.407	0.255	0.226	0.300	0.424	0.424	0.296
0.25	**0.675**	*0.147*	0.663	0.366	0.248	0.219	0.288	0.376	0.389	0.278
0.3	0.588	*0.107*	**0.591**	0.344	0.237	0.203	0.270	0.347	0.357	0.260
						Recall				
0.05	0.329	**0.616**	0.314	0.14	0.092	*0.081*	0.129	0.147	0.136	0.110
0.1	0.371	**0.702**	0.349	0.171	0.114	*0.092*	0149	0.184	0.173	0.127
0.15	0.39	**0.763**	0.382	0.200	0.127	*0.103*	0.158	0.213	0.208	0.140
0.2	0.419	**0.816**	0.401	0.217	0.136	*0.121*	0.160	0.226	0.226	0.158
0.25	0.452	**0.871**	0.445	0.246	0.173	*0.147*	0.193	0.252	0.261	0.186
0.3	0.5	**0.910**	0.504	0.294	0.202	*0.173*	0.230	0.296	0.305	0.221
						F-measure				
0.05	0.495	**0.573**	0.474	0.212	0.139	*0.123*	0.196	0.222	0.206	0.166
0.1	**0.531**	0.498	0.498	0.244	0.163	*0.131*	0.213	0.263	0.247	0.182
0.15	**0.535**	0.412	0.520	0.272	0.173	*0.141*	0.215	0.290	0.284	0.191
0.2	**0.546**	0.328	0.524	0.283	0.177	*0.157*	0.209	0.295	0.295	0.206
0.25	**0.541**	0.251	0.533	0.294	0.207	*0.176*	0.231	0.302	0.213	0.223
0.3	0.540	0.191	**0.544**	0.317	0.218	*0.187*	0.249	0.320	0.329	0.239
Time	20.28s	16.18h	3.91s	4.04s	25.36s	28.89s	27.54s	25.08s	29.3s	3.67m

combined the two keyword sets together, and processed both documents and keywords using the same steps outlined in Section 1.4.1. After excluding papers with less than 11 keywords, we are left with 239 documents. In line with previous experiments, we randomly select five keywords from the ground truth set to serve as observed keywords. Due to the computationally intensive nature of the Gibbs sampling scheme used in BSS, we did not include BSS in this comparison study. When measuring the performance for unsupervised methods, we force observed keywords to be part of the final identified keywords. The precisions, recalls, and F-measures are summarized in Table 1.5. VBSS demonstrates improved performance on all three metrics for all values of γ (except $\gamma = 0.05$). Following VBSS, SS and KPMiner exhibit the next best performance among the evaluated methods. When $\gamma = 0.05$, SS and KPMiner have slightly higher recalls and F-measures. While other methods experience some improvement in their performances with increasing γ values, VBSS consistently demonstrates superior performance over a range of γ values.

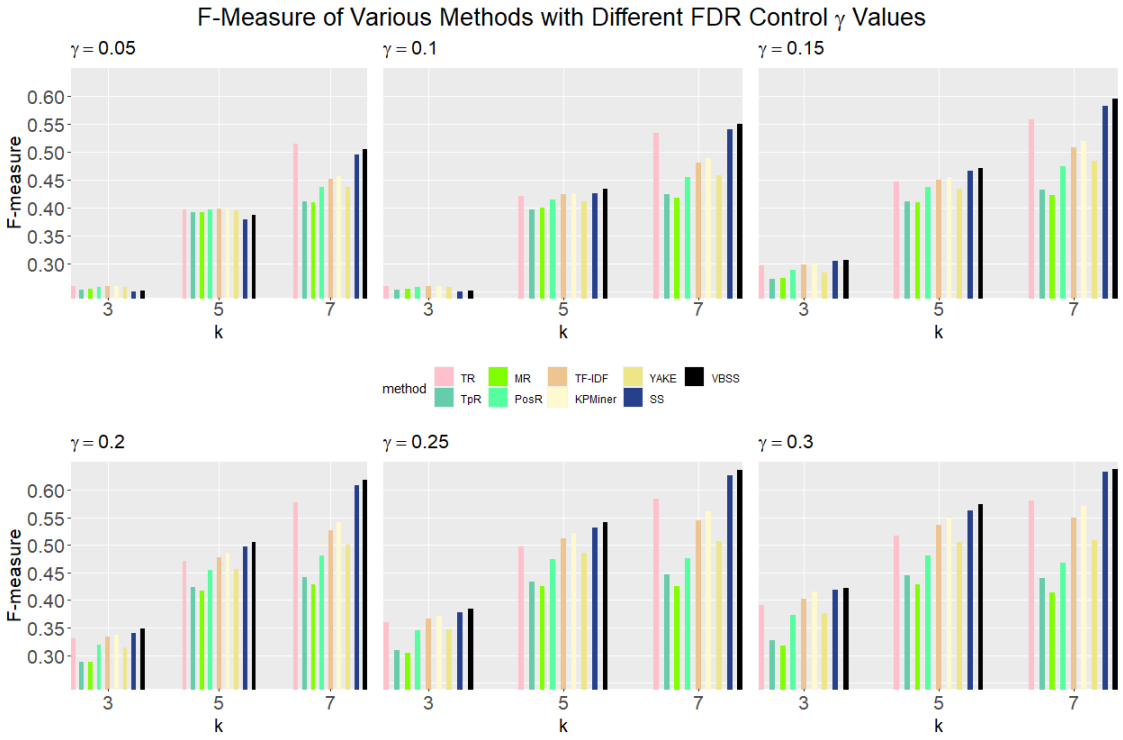
Using the SemEval2010 dataset, we explore how the number of observed keywords can potentially impact the performance of VBSS and other approaches for keyword extraction. For each article, we randomly select $k = 3, 5, 7$ words from the ground truth list of keywords as the observed ones. The overall F-measure for each method v.s. number of observed keywords k is shown in Figure 1.2. As expected, the overall performance of all methods improves as k increases from 3 to 7, since more information becomes available (recall that for unsupervised approaches, we force all observed keywords to be positive). Specifically, when $k = 3$, the overall F-measures of the various methods are more comparable. However, as k increases to 7, VBSS demonstrates a significant improvement over the least performing methods, surpassing the other methods for $\gamma > 0.05$.

We also investigate the impact of the proportion of keywords. We divide the 239 articles into four groups based on the quartiles of proportions of keywords. Group A contains documents with the smallest proportion of keywords (less than or equal to the first quartile),

Table 1.5: SemEval2010 data: the comparison of precisions, recalls and F-measures of various keyword extraction approaches with different FDR control γ.

FDR threshold γ					**Precision**				
	VBSS	SS	TR	TpR	MR	PosR	TF-IDF	KPMiner	YAKE
0.05	**1**	0.984	0.975	*0.961*	0.962	0.973	0.980	0.981	0.968
0.1	**0.979**	0.957	0.936	*0.884*	0.891	0.921	0.940	0.945	0.914
0.15	**0.939**	0.928	0.889	0.817	*0.814*	0.868	0.895	0.904	0.864
0.2	**0.908**	0.891	0.844	0.758	*0.746*	0.814	0.856	0.868	0.818
0.25	**0.835**	0.821	0.766	0.669	*0.654*	0.731	0.787	0.804	0.749
0.3	**0.749**	0.737	0.677	0.583	*0.562*	0.628	0.702	0.726	0.667
					Recall				
0.05	*0.24*	**0.252**	0.25	0.246	0.246	0.249	0.251	0.251	0.248
0.1	**0.28**	0.278	0.272	*0.257*	0.259	0.268	0.273	0.275	0.266
0.15	**0.315**	0.312	0.299	0.275	*0.274*	0.292	0.301	0.304	0.291
0.2	**0.351**	0.345	0.327	0.294	*0.289*	0.315	0.332	0.336	0.317
0.25	**0.401**	0.394	0.368	0.321	*0.314*	0.351	0.378	0.386	0.36
0.3	**0.463**	0.456	0.418	0.361	*0.346*	0.388	0.432	0.444	0.410
					F-measure				
0.05	*0.387*	**0.401**	0.398	0.392	0.392	0.397	0.399	0.400	0.395
0.1	**0.434**	0.431	0.422	*0.398*	0.401	0.415	0.424	0.426	0.412
0.15	**0.472**	0.467	0.448	0.411	*0.410*	0.437	0.451	0.455	0.435
0.2	**0.506**	0.498	0.471	0.423	*0.417*	0.455	0.478	0.485	0.457
0.25	**0.541**	0.533	0.497	0.434	*0.425*	0.475	0.511	0.522	0.486
0.3	**0.573**	0.563	0.517	0.446	*0.429*	0.480	0.536	0.549	0.507

Figure 1.2: SemEval2010 data: the F-measures v.s. different number of observed keywords for various keyword extraction approaches with different FDR control γ.



and Group D consists of articles with the largest proportion of keywords (larger than or equal to the third quartile). For each article, we fix the number of observed keywords to be 5. Consequently, as the proportion of keywords increases, more information is concealed within the documents.

The relationship between the F-measure and the proportion of keywords for different methods, using various γ values, is depicted in Figure 1.3. Due to space limit, we only include SS, TR and KPMiner in Figure 1.3 as they constantly outperform other approaches in our experiments. Again, we control the number of positives identified by SS, TR, and KPMiner based on the results obtained from VBSS and force the observed keywords to be positives for the unsupervised methods TR and KPMiner. The results clearly demonstrate a consistent downward trend in the F-measure as the proportion of keywords increases, regardless of the specific γ and the method employed. Notably, as the FDR cutoff γ becomes larger, this decreasing trend becomes less pronounced, indicating the methods are able to pick up

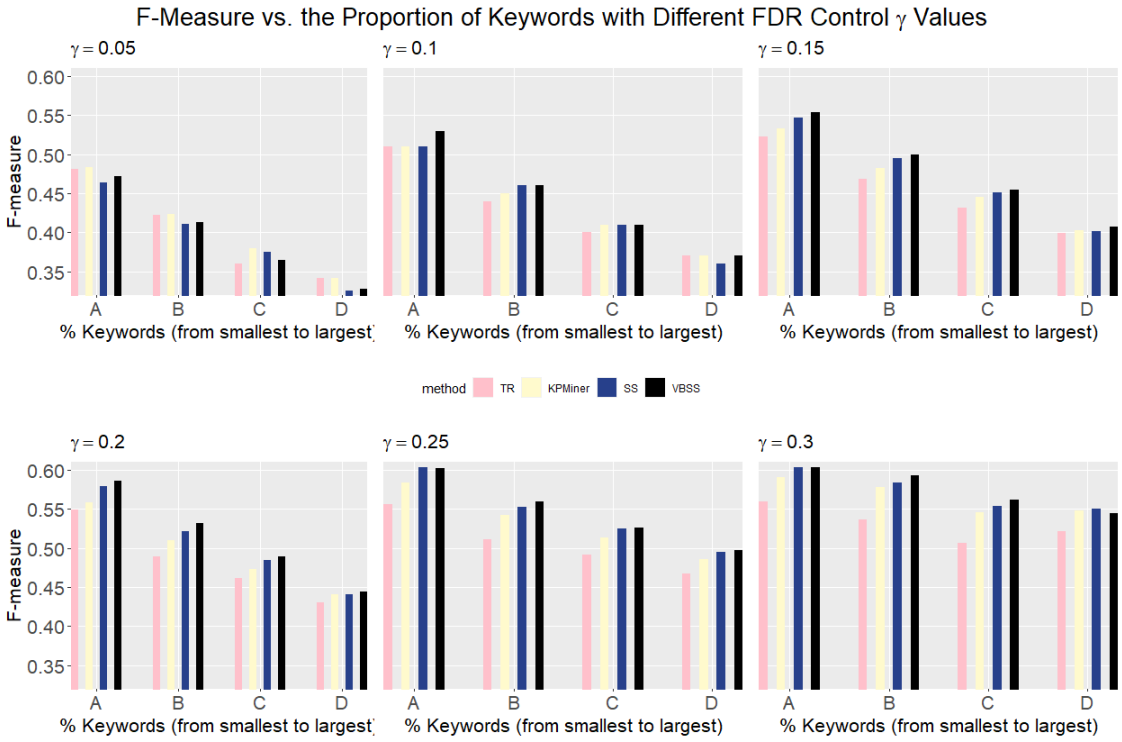
important information eventually regardless of how much is known. When γ is as small as 0.05, SS, TR and KPMiner exhibit a tendency to outperform VBSS (partly due to rounding issues). However, the performance of VBSS is enhanced as γ increases and is the best among all the approaches being compared.

1.4.3. Hulth data

The Hulth dataset [2] is another widely used benchmark dataset comprising over 2,000 abstracts from computer science papers published between 1998 and 2002. Following the preprocessing steps in Wang et al. [1], after removing texts with fewer than 11 keywords, we are left with 1,459 abstracts. It is important to note that these abstracts are considerably shorter in length compared to the documents in the PubMed and SemEval2010 datasets. Our primary focus for the Hulth dataset is to evaluate computational efficiency of VBSS vs. BSS. We do not aim to beat BSS which has demonstrated high accuracy of keyword identification for short documents. Instead, we illustrate the substantial improvement in computation efficiency. In Figure 1.4, we present how much time VBSS and BSS take as the number of candidate words change. In the left figure, the time of both methods exhibits an approximate quadratic relationship to the number of candidate words n, although the trend with VBSS is harder to identify as it takes much shorter time than BSS. The box-plots on the right clearly demonstrate how faster VBSS is than BSS. Overall, it takes VBSS 1.93 minutes to complete the identification across 1,459 abstracts while BSS spends 47.2 hours. The comparison highlights the significant improvement in terms of the computation efficiency using variational inference, thus making VBSS an attractive option for keyword identification tasks.

The precisions, recalls, and F-measures of VBSS and BSS across different FDR cutoff values are displayed in Table 1.2 in Appendix A. VBSS exhibits higher precisions for smaller γ's. Meanwhile, BSS consistently achieves better recalls regardless of the chosen γ values.

Figure 1.3: SemEval2010 data: the relationship between the F-measure and the proportion of keywords for various keyword extraction approaches with different FDR control γ.

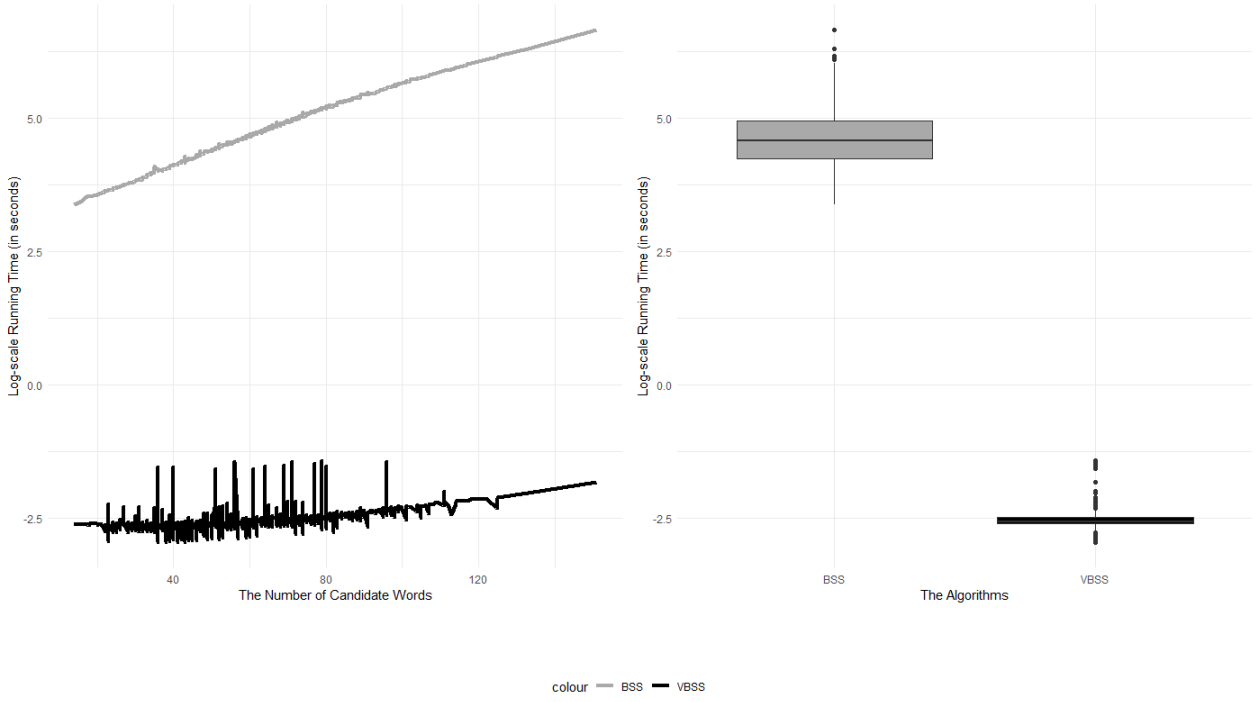


For the overall performance (F-measure), VBSS performs better for larger γ's while BSS is better for smaller γ's.

1.5. Discussion and Future Direction

The keyword identification problem has attracted significant attention and research efforts, yet the semi-supervised setting, which assumes a subset of keywords is known, remains under-explored. Recently, Wang et al. [1] proposed a semi-supervised Bayesian keyword identification approach which shows superior performance on short articles. However, the computational burden of their proposed method has prevented it from being applied effectively on longer articles. To address this challenge, we propose a novel method called variational Bayesian semi-supervised keyword extraction (VBSS). Our approach employs variational inference to approximate the joint posterior distribution, leading to a significant reduction in computational time, thereby enabling its application to longer articles. In ad-

Figure 1.4: Hulth data: the plot of running time (in seconds) of VBSS and BSS versus the number of candidate words n (left) and the box-plots of the time consumption of VBSS and BSS. BSS is visualized in grey and VBSS is in black.



dition, we introduce additional parameters to allow greater model flexibility. As a result, on long articles, VBSS exhibits remarkable performance particularly with larger γ values in comparison to a broad spectrum of existing methods. In addition, although VBSS is developed mainly for long articles, our method still exhibits impressive computational efficiency in short articles while preserving competitive performance.

For practical implementation of our VBSS method, the users can specify the FDR value γ to select the number of identified keywords. We suggest using a relatively larger γ as it is usually associated with better performance as shown in real-world examples. In our preliminary experiment (results not shown for conciseness), we notice that the choice of priors on $\boldsymbol{\theta}$ has an effect on the results. For a particular dataset, the user might want to try different priors such as the prior used in Wang et al. [1] and other reasonable alternatives that carry information about the words importance.

We point out several directions for future research. In our model, to ensure an analytical form of the solution in each optimization step of our variational inference, we have assumed a probit model for the keyword probability and introduce additional latent variables to augment the model. A possible alternative to our model choice is to use the well-known logit link function instead. Similar to our data augmentation idea, to circumvent the need of numerical integration, fast computation could potentially be achieved through the idea of Pólya-Gamma augmentation [28]. In this paper, as a proof of the concept, we use the mean-field variational inference which assumes independence among the parameters. We adopt the coordinate ascent algorithm to optimize the variational objective function. Another choice is stochastic variational inference, whose objective function is optimized through stochastic optimization with noisy natural gradients [29]. In addition, one could consider a more relaxed group of variational distributions that accounts for possible dependencies among the parameters, for example, through some reparameterization scheme [30–33]. Furthermore,

alternative divergence beyond $\mathcal{KL}$-divergence, such as f-divergence [34], might be considered for tighter bounds [35].

Chapter 2

VARIATIONAL BAYESIAN MULTIMODAL MULTIPLE INSTANCE CLASSIFICATION

2.1. Introduction

Multiple instance learning (MIL), a machine learning paradigm initially proposed by Dietterich et al. [36], has attracted numerous research efforts due to its natural fit in various real-world scenarios such as medical imaging [37], computer vision [38], document classification [39], etc. In the setting of MIL, the label or value of a single object (called bag), is determined by a collection of feature vectors (called instances) contained in this bag, instead of by an individual feature vector. These instances share the same set of attributes. For example, the activity of a drug is determined by multiple molecular conformations in this drug [36], and the overall sentiment of a text could be classified based on sentences or elementary discourse units [40]. The supervision is provided by the label or value of the entire bag, instead of instances, which leads to a weakly supervised learning process. Determined by the specific problem setting, the objective of MIL can be adjusted into binary [36] or multiple-class classification [41], regression [42, 43], ranking [44] or clustering [45].

The majority of MIL work focuses on binary multiple instance classification (MIC), where the bag label is either positive or negative. According to Amores [46], the binary MIC algorithms can be further divided into three categories based on the means of data extraction and exploitation: instance-space (IS), bag-space (BS), and embedded-space (ES). MIC aims to learn a function to quantify the probability of a bag being positive. IS methods train an instance level classifier then combine to create a bag level classifier. BS and ES approaches treat each bag as a whole entity and train the bag-level classifier using the global informa-

tion. BS approaches work through the distance between each pair of bags then consequently employ distance-based classifier such as K-Nearest Neighbor, or similarly into any kernel-based classifier such as Support Vector Machine (SVM). ES approaches, on the other hand, map the whole bag into a single vector by summarizing the relevant information available in the bag. We refer readers to Amores [46], Carbonneau et al. [47] and Xiong et al. [48] for more detailed and structured reviews of existing MIL methods. We note that existing MIL methods, primarily rooted in computer science, often prioritize algorithms over explicit data models. This often limits interpretability, making it difficult to understand how these methods arrive at their predictions. To our knowledge, only MILR [49] and MICProB [50] leverage statistical models to provide interpretable results. MILR is an IS method, which employs a logistic regression model with an optional Lasso penalty term at the instance level and then associates the bag probability to the predicted instance probabilities via the standard multi-instance assumption (i.e., a positive bag has at least one positive instance and a negative bag only has negative instances). By contrast, MICProB does not belong to any of the three categories (IS, BS or ES). It adopts the so-called PI framework [51], which assumes that only a subset of instances, called primary instances (PIs), are relevant to bag labels, while the rest are considered irrelevant. MICProB is based on a unique two-tier probit regression model, whose inner probit regression focuses on identifying primary instances, and the outer probit regression focuses on predicting bag labels. The transparent structure of MICProB, combined with its Bayesian hierarchical setup, allows for interpretability, statistical inference and uncertainty quantification at both bag and instance levels. It has been demonstrated in Xiong et al. [50] that MICProB exhibits better performance than existing MIC algorithms in both synthetic data and certain real-world applications. However, MICProB's MCMC sampling makes it computationally expensive for large datasets, a limitation shared by many other MIC methods [46, 47, 52].

In recent years, the study of information from multiple channels has gained increasing attention across various domains, driven by the recognition that multi-source data offers

complementary perspectives. For example, Ma et al. [53] studied the effects of textual content and user-provided photos on hotel review helpfulness. Similarly, Al-Tameemi et al. [54] proposed a sentiment classification model that integrates textual and visual data. Xu et al. [55] integrated visual clues from lip movements to augment the accuracy of audio speech recognition systems. Tang et al. [56] introduced a multi-task deep neural network to analyze single-cell multi-modality data such as spatial transcriptomics and gene expressions. Li et al. [57] conducted a comprehensive review of various medical imaging techniques for cardiovascular diseases, which utilize images from diverse sources such as PET, CT, and MRI scans. The contributions of information provided through different modalities are usually unbalanced. For instance, Ma et al. [53] pointed out that images do not convey sufficient clues compared to texts. Also, Xu et al. [55] regarded visual data as complementary rather than primary information in audio speech recognition tasks.

Adapting MIC to multimodal instances presents an intriguing avenue, given its cost-effective weakly-supervised labeling. However, compared to unimodal MIC, the area of multimodal MIC where instances from different views contribute to the same bag is severely under-developed. Existing multimodal MIC methods are primarily found in medical analysis, often with highly specialized input requirements. For instance, in Sahasrabudhe et al. [58], one modality comprises blood cell images, while the other modality consists of two single values: patient age and lymphocyte count measured in cells per liter of blood. Li et al. [59] uses both fundus photos and OCT scans to diagnose retinal disease. Thus, these approaches are tailored to specific tasks and do not address the broader challenges of general multimodal instances.

Built upon MICProB, we aim to develop a novel Bayesian method called Variational Multimodal Multiple Instance Classification (vMMIC). This is the first attempt to adapt MIL into multimodal multi-instance scenarios through a rigorous Bayesian hierarchical model, suitable for the general bag-instance structured inputs. Our modeling decisions are made

to boost algorithm scalability without imposing limiting assumptions or sacrificing interpretability. Without loss of generality, we focus on bimodal instances. For example, consider a Yelp review containing 10 text sentences and 4 images. In this scenario, the review has 10 instances for the first modality (text) and 4 for the second (image). To significantly improve computational efficiency, we leverage mean-field variational inference (MFVI) for approximating the joint posterior distribution. This approach, combined with data augmentation techniques, allows vMMIC to handle large-scale datasets effectively. Furthermore, in vMMIC, each instance is assigned a binary indicator with "1" indicating a primary instance and "0" otherwise. Primary instances in each bag collectively contribute to the probability of a bag being positive in a linear manner, through a probit link. This setup would allow us to identify "responsible" instances (e.g., finding which sentences and or images contribute to a positive review). Besides, benefitting from its Bayesian framework, vMMIC further offers statistical inference and uncertainty quantification for any model parameter or prediction, a feature often overlooked in existing algorithm-driven methods.

The remainder of this paper is organized as follows. Section 2.2 introduces the construction of the Bayesian model and the detailed description of vMMIC algorithm, including analytical derivation for variational inference. We compare our vMMIC and existing methods in simulation experiments in Section 2.3. In Section 2.4, we apply our vMMIC to a Yelp dataset which consists of restaurant reviews embedded with user-provided images. Section 2.5 summarizes our work and discusses future directions.

2.2. Model and Algorithm

For illustration purpose, we focus on a dual-modality framework. However, what's described below can be easily extended to data with more than two modalities. We assume the dataset comprises n independent bags, and the i-th bag, denoted by B_i, has a single binary response $y_i \in (0, 1)$, $i \in (1, \cdots, n)$. The bag B_i has m_i instances where the first m_i^0 instances belong to the first modality and the other m_i^1 instances for the second modality, thus

$m_i = m_i^0 + m_i^1$. Instances from the first modality are described by d^0 features and instances from the second modality are described by d^1 features. We set indicator $w_{ij} = 0$ if the j-th instance is from the first modality and $w_{ij} = 1$ if otherwise. Therefore, $\sum_j (1 - w_{ij}) = m_i^0$ and $\sum_j w_{ij} = m_i^1$. Let $\boldsymbol{x}_{ij} = (x_{ij1}, \ldots, x_{ijd^0})$ if $w_{ij} = 0$. For all instances from the second modality (i.e., $w_{ij} = 1$), their $\boldsymbol{x}_{ij}$ is undefined and for simplicity, we set $\boldsymbol{x}_{ij} \equiv \mathbf{0}$ without affecting computation results. Similarly, let $\boldsymbol{z}_{ij} = (z_{ij1}, \ldots, z_{ijd^1})$ if $w_{ij} = 1$ and $\boldsymbol{z}_{ij} \equiv \mathbf{0}$ otherwise. Therefore, one bag can be written as $\left\{ (\boldsymbol{x}_{i1}, \cdots, \boldsymbol{x}_{im_i})^T, (\boldsymbol{z}_{i1}, \cdots, \boldsymbol{z}_{im_i})^T, (w_{i1}, \cdots, w_{im_i})^T, y_i \right\}$. By stacking $\boldsymbol{x}_{ij}$ row-wisely, we obtain an m_i by d^0 matrix, denoted by $\boldsymbol{X}_i$, for the first modality. Similarly, we obtain an m_i by d^1 matrix, denoted by $\boldsymbol{Z}_i$, for the second modality. We assume that only part of the instances in a bag will be informative and contribute to the bag label. We call those instances the primary instances. We use a latent indicator $\delta_{ij} = 1$ to represent that the j-th instance is a primary instance in the i-th bag and $\delta_{ij} = 0$ otherwise.

2.2.1. Model and priors specification

At the instance level, we use a probit regression to model the primary indicator δ_{ij}. Namely,

$$\delta_{ij} = \text{sign}\left(U_{ij}\right),$$
$$U_{ij} \sim N\left((1 - w_{ij})\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right), 1\right),$$

where a and c are two intercepts, $\boldsymbol{b} = (b_1, \cdots, b_{d_0})^T$ is a column vector of regression coefficients describing the effects of covariates from the first modality on the primary instance indicator δ_{ij} and $\boldsymbol{d} = (d_1, \cdots, d_{d_1})^T$ is a column vector of regression coefficients describing the effects of covariates from the second modality on δ_{ij}. Thus, $P\left(\delta_{ij} = 1 \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right) = \Phi\left((1 - w_{ij})\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)$.

At the bag level, we adopt another probit link to connect the bag label with instances $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. That is,

$$
\begin{aligned}
y_i &= \operatorname{sign}\left(y_i^*\right), \\
y_i^* &\sim N\left(\alpha + \sum_{j=1}^{m_i} \delta_{ij}\left(1 - w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta} + \sum_{j=1}^{m_i} \delta_{ij} w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma},\ 1\right),
\end{aligned}
$$

where α is the intercept, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_{d_0})^T$ is a column vector of regression coefficients for the covariates from the first modality, $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_{d_1})^T$ is a column vector of regression coefficients for the second modality. Thus, $P(y_i = 1 \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i) = \Phi(\alpha + \sum_{j=1}^{m_i} \delta_{ij}\left(1 - w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta} + \sum_{j=1}^{m_i} \delta_{ij} w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma})$. An instance only contributes to the bag label if the corresponding indicator variable $\delta_{ij} = 1$.

In our model, y_i^* and U_{ij} are the latent variables introduced to guarantee the posterior distributions for all the parameters are of closed forms [20]. Let $\boldsymbol{y} = (y_1, \cdots, y_n)^T$ be a column vector of length n containing all bag labels and $\boldsymbol{y}^* = (y_1^*, \cdots, y_n^*)^T$ be associated latent variables. Let $\boldsymbol{\Delta}$ and $\boldsymbol{U}$ be column vectors of length $\sum_{i=1}^{n} m_i$ containing all primary instances indicators and associated latent variables. Let $\boldsymbol{\Theta} = (\boldsymbol{y}^*, \boldsymbol{\Delta}, \boldsymbol{U}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, a, \boldsymbol{b}, c, \boldsymbol{d})$ include all the parameters and the latent variables in the model. We set conjugate priors of α, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, a, $\boldsymbol{b}$, c and $\boldsymbol{d}$ as $N\left(\mu_\alpha, \sigma_\alpha^2\right)$, $N\left(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right)$, $N\left(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}\right)$, $N\left(\mu_a, \sigma_a^2\right)$, $N\left(\boldsymbol{\mu}_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}\right)$, $N\left(\mu_c, \sigma_c^2\right)$, and $N\left(\boldsymbol{\mu}_{\boldsymbol{d}}, \Sigma_{\boldsymbol{d}}\right)$, respectively. Note that μ_α, σ_α^2, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$, $\Sigma_{\boldsymbol{\beta}}$, $\boldsymbol{\mu}_{\boldsymbol{\gamma}}$, $\Sigma_{\boldsymbol{\gamma}}$, μ_a, σ_a^2, $\boldsymbol{\mu}_{\boldsymbol{b}}$, $\Sigma_{\boldsymbol{b}}$, μ_c, σ_c^2, $\boldsymbol{\mu}_{\boldsymbol{d}}$ and $\Sigma_{\boldsymbol{d}}$ are user-defined hyperparameters. We suggest using diffuse or vague priors. For all of our data analyses, we set prior means as zero, prior variances of each intercept as 16, and prior covariances of slopes as diagonal matrices with diagonals being 4, as suggested by Xiong et al. [50]. The hierarchical structure of the proposed model is

illustrated in Figure 2.1. The full probability model is given by

$$
\begin{aligned}
p(\boldsymbol{y}, \boldsymbol{\Theta} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}) = & \, p\left(\boldsymbol{y} \mid \boldsymbol{y}^*\right) \times p(\boldsymbol{y}^* \mid \alpha, \boldsymbol{\Delta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}) \\
& \times p\left(\boldsymbol{\Delta} \mid \boldsymbol{U}\right) \times p\left(\boldsymbol{U} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right) \\
& \times p\left(\alpha \mid \mu_\alpha, \sigma_\alpha^2\right) \times p\left(\boldsymbol{\beta} \mid \boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right) \times p\left(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}\right) \\
& \times p\left(a \mid \mu_a, \sigma_a^2\right) \times p\left(\boldsymbol{b} \mid \boldsymbol{\mu}_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}\right) \times p\left(c \mid \mu_c, \sigma_c^2\right) \times p\left(\boldsymbol{d} \mid \boldsymbol{\mu}_{\boldsymbol{d}}, \Sigma_{\boldsymbol{d}}\right) \\
= & \prod_{i=1}^{n} p\left(y_i \mid y_i^*\right) \times p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \\
& \times \prod_{i=1}^{n} \prod_{j=1}^{m_i} p\left(\delta_{ij} \mid U_{ij}\right) \times p\left(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right) \\
& \times p\left(\alpha \mid \mu_\alpha, \sigma_\alpha^2\right) \times p\left(\boldsymbol{\beta} \mid \boldsymbol{\mu}_{\boldsymbol{\beta}}, \Sigma_{\boldsymbol{\beta}}\right) \times p\left(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \Sigma_{\boldsymbol{\gamma}}\right) \\
& \times p\left(a \mid \mu_a, \sigma_a^2\right) \times p\left(\boldsymbol{b} \mid \boldsymbol{\mu}_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}\right) \times p\left(c \mid \mu_c, \sigma_c^2\right) \times p\left(\boldsymbol{d} \mid \boldsymbol{\mu}_{\boldsymbol{d}}, \Sigma_{\boldsymbol{d}}\right).
\end{aligned}
$$

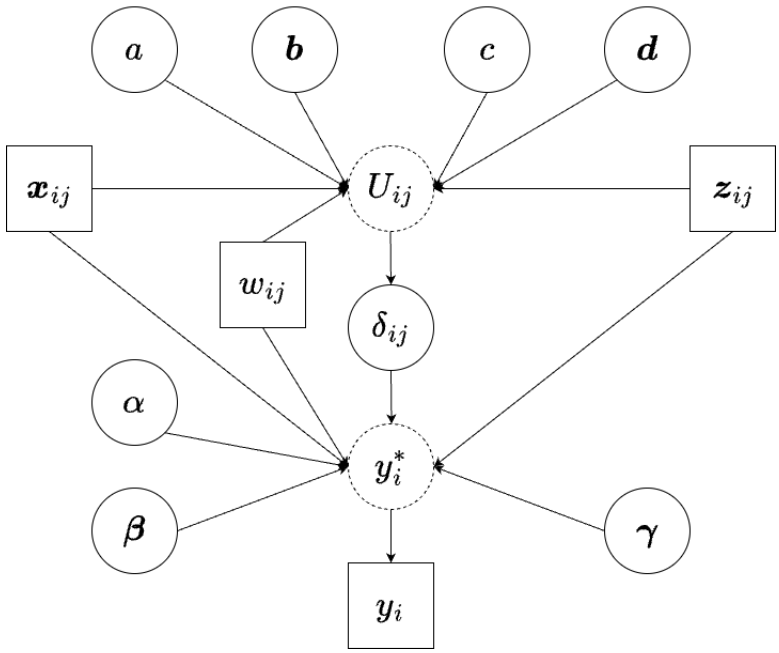
2.2.2. Mean-field variational inference

We adopt the variational inference machinery to approximate the posterior distribution via minimizing the $\mathcal{KL}$-divergence between the candidate approximation and the exact true posterior [19, 60]. Specifically, we follow the mean-field variational inference assumption where the variational distributions of different parameters are mutually independent. In our context, MFVI aims to find a group of independent variational distributions q's such that

$$
\begin{aligned}
p\left(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right) \approx & \left(\prod_{i=1}^{n} q\left(y_i^*\right) \prod_{j=1}^{m_i} q\left(\delta_{ij}\right) q\left(U_{ij}\right)\right) \\
& \times q\left(\alpha\right) q\left(\boldsymbol{\beta}\right) q\left(\boldsymbol{\gamma}\right) q\left(a\right) q\left(\boldsymbol{b}\right) q\left(c\right) q\left(\boldsymbol{d}\right).
\end{aligned}
$$

The minimization of $\mathcal{KL}$-divergence can be alternatively achieved by maximizing a constant minus $\mathcal{KL}$-divergence, which is known as the evidence lower bound (ELBO) [18, 61]. We use the coordinate ascent algorithm to maximize the ELBO by iteratively up-

Figure 2.1: Bayesian hierarchical model structure of vMMIC. Observed data, including instances $\boldsymbol{x}_{ij}$, $\boldsymbol{z}_{ij}$, indicator w_{ij} and bag label y_i, are showed in square boxes. Latent variables introduced for computation purpose, including y_i^* and U_{ij}, are showed in dashed circles.



dating the variational distribution of one parameter while holding all other variational distributions fixed. It is easy to show that the optimal variational distribution $q^*(\theta_i) \propto \exp\left(\mathbb{E}_{\boldsymbol{\Theta}_{-i}}\left[\log\left(p\left(\theta_i \mid \boldsymbol{\Theta}_{-i}, \boldsymbol{y}\right)\right)\right]\right)$ in each iteration, where θ_i denote the ith element of $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}_{-i}$ denotes $\boldsymbol{\Theta}$ but excluding θ_i. Following this, we derive the optimal variational distributions of parameters in Θ, respectively, of which the details are provided in Section B.1 in Appendix B. The optimized variational distribution of a particular parameter depends on the expectations of other parameters specified below, which can be directly calculated from the variational distributions corresponding to each individual parameter:

- $\mathbb{E}_{q^*(\alpha)}[\alpha] = \sigma_\alpha^2/\left(1+n\sigma_\alpha^2\right)\left(\frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i=1}^n\left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] - \mathbb{E}_{q^*(\boldsymbol{l}_i')}\left[\boldsymbol{l}_i'\right]^T \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)\right)$.
- $\mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] = \boldsymbol{V}_{\boldsymbol{\beta}}\left(\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^n\left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right)\boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}[\boldsymbol{l}_i] - \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i'^T\right] \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)$, where $\boldsymbol{V}_{\boldsymbol{\beta}} = \left(\Sigma_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^n \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i^T\right] \boldsymbol{X}_i\right)^{-1}$, $l_i = \left(\delta_{i1}\left(1-w_{i1}\right), \cdots, \delta_{im_i}\left(1-w_{im_i}\right)\right)^T$, and $l_i' = \left(\delta_{i1}w_{i1}, \cdots, \delta_{im_i}w_{im_i}\right)^T$.

- $\mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}] = \boldsymbol{V}_{\boldsymbol{\gamma}}\left(\Sigma_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{i=1}^{n}\left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right)\boldsymbol{Z}_i^T\mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i'\right] - \boldsymbol{Z}_i^T\mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i'\boldsymbol{l}_i'^T\right]\boldsymbol{X}_i\mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}]\right)$, where $\boldsymbol{V}_{\boldsymbol{\gamma}} = \left(\Sigma_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^{n}\boldsymbol{Z}_i^T\mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i'\boldsymbol{l}_i'^T\right]\boldsymbol{Z}_i\right)^{-1}$.

- $\mathbb{E}_{q^*(a)}[a] = \left(\frac{\mu_a}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-w_{ij})\left(\mathbb{E}_{q^*(U_{ij})}[U_{ij}] - \boldsymbol{x}_{ij}\mathbb{E}_{q^*(\boldsymbol{b})}[\boldsymbol{b}]\right)\right) / \left(\frac{1}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-w_{ij})\right)$.

- $\mathbb{E}_{q^*(\boldsymbol{b})}[\boldsymbol{b}] = \boldsymbol{V}_{\boldsymbol{b}}\left(\Sigma_{\boldsymbol{b}}^{-1}\boldsymbol{\mu}_{\boldsymbol{b}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}^T(1-w_{ij})\left(\mathbb{E}_{q^*(U_{ij})}[U_{ij}] - \mathbb{E}_{q^*(a)}[a]\right)\right)$, where $\boldsymbol{V}_{\boldsymbol{b}} = \left(\Sigma_{\boldsymbol{b}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}(1-w_{ij})\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ij}\right)^{-1}$.

- $\mathbb{E}_{q^*(c)}[c] = \left(\frac{\mu_c}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}w_{ij}\cdot\left(\mathbb{E}_{q^*(U_{ij})}[U_{ij}] - \boldsymbol{z}_{ij}\mathbb{E}_{q^*(\boldsymbol{d})}[\boldsymbol{d}]\right)\right) / \left(\frac{1}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}w_{ij}\right)$.

- $\mathbb{E}_{q^*(\boldsymbol{d})}[\boldsymbol{d}] = \boldsymbol{V}_{\boldsymbol{d}}\left(\Sigma_{\boldsymbol{d}}^{-1}\boldsymbol{\mu}_{\boldsymbol{d}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{z}_{ij}^{\top}w_{ij}\left(\mathbb{E}_{q^*(U_{ij})}[U_{ij}] - \mathbb{E}_{q^*(c)}[c]\right)\right)$, where $\boldsymbol{V}_{\boldsymbol{d}} = \left(\Sigma_{\boldsymbol{d}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}w_{ij}\boldsymbol{z}_{ij}^{\top}\boldsymbol{z}_{ij}\right)^{-1}$.

- If $y_i = 1$, $\mathbb{E}_{q^*(y_i^*)}[y_i^*] = m_{y_i^*} + \phi(-m_{y_i^*})/\left[1-\Phi\left(-m_{y_i^*}\right)\right]$, where $m_{y_i^*} = \mathbb{E}_{q^*(\alpha)}[\alpha] + \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]\boldsymbol{X}_i\mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] + \mathbb{E}_{q^*(\boldsymbol{l}_i')}\left[\boldsymbol{l}_i'\right]\boldsymbol{Z}_i\mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]$; If $y_i = 0$, $\mathbb{E}_{q^*(y_i^*)}[y_i^*] = m_{y_i^*} - \phi(-m_{y_i^*})/\Phi\left(-m_{y_i^*}\right)$.

- $\mathbb{E}_{q^*(\delta_{ij})}[\delta_{ij}] = A/(A+B)$, where

  $$A = A^*\left(1-\Phi\left(-m_{U_{ij}}\right)\right) \times \exp\left\{\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(y_i^* - \alpha - \sum_{j'\neq j}^{m_i}\delta_{ij'}(1-w_{ij'})\boldsymbol{x}_{ij'}\boldsymbol{\beta} - \sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\left((1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta} + w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\right)\right]\right\} \times \exp\left\{-\frac{1}{2}\mathbb{E}_{q^*(-\delta_{ij})}\left[(1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^T + w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{z}_{ij}^T\right]\right\},$$

  $B = B^*\cdot\Phi\left(-m_{U_{ij}}\right)$, $m_{U_{ij}} = \mathbb{E}_{q^*(a)}[a] + \boldsymbol{x}_{ij}\mathbb{E}_{q^*(\boldsymbol{b})}[\boldsymbol{b}]$, A^* and B^* are A and B computed in the previous iteration.

- $\mathbb{E}_{q^*(U_{ij})}[U_{ij}] = \dfrac{(B-A)\cdot m_{U_{ij}}\cdot\Phi\left(-m_{U_{ij}}\right) + A\cdot m_{U_{ij}} + \frac{A-B}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}m_{U_{ij}}^2\right\}}{A + (B-A)\cdot\Phi\left(-m_{U_{ij}}\right)}$.

Therefore, instead of iteratively sampling from the joint posterior distribution, we directly estimate parameters for the posterior. In the next subsections, we provide details of parameter estimation and prediction for new bags.

2.2.3. Parameter estimation and new bag prediction

To estimate the parameters of those variational distributions, we initialize the parameters of vMMIC with random values and update the values using coordinate ascent until convergence. As detailed in Section 2.2.2, each variational distribution has a closed form, facilitating straightforward updates to distributional parameters, such as the means and variances. Convergence is monitored at each iteration. The probability of being a positive bag for the i-th bag in the h-th iteration is estimated as $\hat{p}_i^{(h)} = \Phi(\alpha^{(h)} + \sum_{j=1}^{m_i} \delta_{ij}^{(h)} (1 - w_{ij}) \boldsymbol{x}_{ij}\boldsymbol{\beta}^{(h)} + \sum_{j=1}^{m_i} \delta_{ij}^{(h)} w_{ij} \boldsymbol{z}_{ij}\boldsymbol{\gamma}^{(h)})$, $h \in (1, \ldots, miter)$, $i \in (1, \ldots, n)$ and $miter$ is the maximum number of iterations allowed. The algorithm terminates when the average difference between two consecutive probability estimates across n training bags is sufficiently small, or when $miter$ is reached. All parameters and latent variables in $\Theta = (\boldsymbol{y}^*, \boldsymbol{\Delta}, \boldsymbol{U}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, a, \boldsymbol{b}, c, \boldsymbol{d})$, are estimated based on the expectations calculated in the final iteration.

Based on these estimated parameters, we can predict the label of a new bag, say $\tilde{B}$, which consists of $\tilde{m}$ instances in total (compromising $\tilde{m}^0$ instances from the first modality and $\tilde{m}^1$ instances from the second modality), as well as identify the primary instances. In the new bag, let instances from the first modality be characterized by $\tilde{\boldsymbol{X}}$, with $\tilde{\boldsymbol{X}} = \{\tilde{\boldsymbol{x}}_1, \cdots, \tilde{\boldsymbol{x}}_{\tilde{m}}\}$, and those from the second by $\tilde{\boldsymbol{Z}}$, with $\tilde{\boldsymbol{Z}} = \{\tilde{\boldsymbol{z}}_1, \cdots, \tilde{\boldsymbol{z}}_{\tilde{m}}\}$. The probability that an instance j is primary can be estimated as $\hat{P}(\tilde{\delta}_j = 1) = \Phi((1 - \tilde{w}_j)(\widehat{a} + \tilde{\boldsymbol{x}}_j\widehat{\boldsymbol{b}}) + \tilde{w}_j(\widehat{c} + \tilde{\boldsymbol{z}}_j\widehat{\boldsymbol{d}}))$, and the probability that the bag is positive can be estimated as $\hat{P}(\tilde{y} = 1) = \Phi(\widehat{\alpha} + \sum_{j=1}^{\tilde{m}} \widehat{\tilde{\delta}}_j (1 - \tilde{w}_j) \tilde{\boldsymbol{x}}_j\widehat{\boldsymbol{\beta}} + \sum_{j=1}^{\tilde{m}} \widehat{\tilde{\delta}}_j \tilde{w}_j \tilde{\boldsymbol{z}}_j\widehat{\boldsymbol{\gamma}})$.

2.3. Simulation Studies

2.3.1. Simulation settings

We evaluate the performance of our vMMIC and other 9 MIC methods across various simulated scenarios. Factors that potentially influence the (relative) performance are varied,

including sample size n, number of instances m, ratio r between the number of instances from the first modality and that from the second modality, as well as the proportions of primary instances within each modality, denoted as $\overline{\text{PPI}}_0$ and $\overline{\text{PPI}}_1$, respectively. We model the modality indicator w_{ij} for the j-th instance in the i-th bag using a Bernoulli distribution with probability $1/(r+1)$. Subsequently, the covariate x_{ijk} is independently generated from $N(0,1)$ if $w_{ij}=0$ and the covariate z_{ijk} is independently generated from $N(-1,1)$ otherwise. The primary status indicator δ_{ij} is generated from another Bernoulli distribution with probability $(1-w_{ij})(a+\boldsymbol{x}_{ij}\boldsymbol{b})+w_{ij}(c+\boldsymbol{z}_{ij}\boldsymbol{d})$, where a and $\boldsymbol{b}$ are the intercept and the slope associated with instances from the first modality, and c and $\boldsymbol{d}$ are the intercept and the slope associated with instances from the second modality. Following this, the bag label y_i is generated from a Bernoulli distribution with probability $\alpha+\sum_{j=1}^{m_i}\delta_{ij}(1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta}+\sum_{j=1}^{m_i}\delta_{ij}w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}$, where α is the intercept, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are slopes associated with instances from the first and the second modalities, respectively.

For simplicity, we assume all bags have the same number of instances, which means $m_i = m$, $\forall i \in (1,\cdots,n)$. We vary $n \in \{250,500,1000,2000,4000,16000\}$, $m \in \{10,20,30,40\}$, $r \in \{1,2,3,4\}$ and $\overline{\text{PPI}}_1 \in \{0.1,0.35,0.6,0.85\}$. The number of covariates is set as 16 for instances from both modality 1 and modality 2. We set $\boldsymbol{\beta}^T = (\underbrace{-1,\cdots,-1}_{8},\underbrace{1,...,1}_{8})$ and $\boldsymbol{\gamma}^T = (\underbrace{-0.5,\cdots,-0.5}_{8},\underbrace{0.5,...,0.5}_{8})$. We set $\boldsymbol{b}^T = (\underbrace{1,\cdots,1}_{16})$ and $\boldsymbol{d}^T = (\underbrace{-0.5,\cdots,-0.5}_{8},\underbrace{0.5,...,0.5}_{8})$ and vary a and c to adjust for specific combination of $\overline{\text{PPI}}_0$ and $\overline{\text{PPI}}_1$. Each factor is varied independently, while the others remain fixed at the basic setting with $n=500$, $m=20$, $r=4$, $\overline{\text{PPI}}_0 = \overline{\text{PPI}}_1 = 0.35$. Under each setting, we generate 50 independent replication datasets and 300 test bags in each replicate. The performance is measured by the averaged area under the Receiver Operating Characteristic curve (AUROC) across 50 replicates.

2.3.2. Methods included in comparison

Xiong et al. [50] examined 16 unimodal MIC methods. Seven of these methods are from the IS category: EMDD [62], MI-SVM [63], mi-SVM [64], MILR [49], SI-SVM [47], SI-kNN [47], and MILBoost [65]. Five methods belong to the BS category: CkNN [66], NSK-SVM [67], EMD-SVM [68], miGraph [69], MInD [70]. Three methods are categorized under the ES category: MILES [71], BoW [46], CCE [72]. MICProB does not fall into any of the three categories. Xiong et al. [50] assessed the performance of these methods across different performance tiers, distinguishing between top, middle, and bottom-performing groups. For clarity, we focus solely on the top and middle performers, which include MICProB, NSK-SVM, EMD-SVM, MInD, MI-SVM, mi-SVM, MILES, MILR, and miGraph. Only MICProB and our vMMIC are capable of identifying primary instances. Importantly, all existing methods compared in our study focus solely on instances from a single modality. Consequently, within each bag, we create a combined instance matrix by arranging instances from modality 1 on the left upper side and instances from modality 2 on the right bottom side, with all other entries filled with zeros. MICProB is implemented using the provided R code from Xiong et al. [50], MILR is implemented using the R package milr [49], and the remaining methods are implemented using the MATLAB MILSurvey toolbox developed by Carbonneau et al. [47]. For vMMIC, we develop our own R code that will be provided via a GitHub link. Default settings are adopted for all methods where applicable. For example, the parameters of MICProB are initialized with random values, and the Gibbs sampler is run for 100,000 iterations, discarding the first half as burn-ins.

In our simulation, certain computationally intensive methods are excluded from settings where their runtime on a single replication dataset exceeds 1.5 hours, using a Windows 10 Operating System with an Intel(R) Core(TM) i5-7400 CPU operating at 3.00GHz and 24.0 GB of RAM. We also note that the multimodal MIC methods that are specifically designed

for certain tasks, as mentioned in the introduction (e.g., Sahasrabudhe et al. [58] and Li et al. [59]), are not included in our comparison.

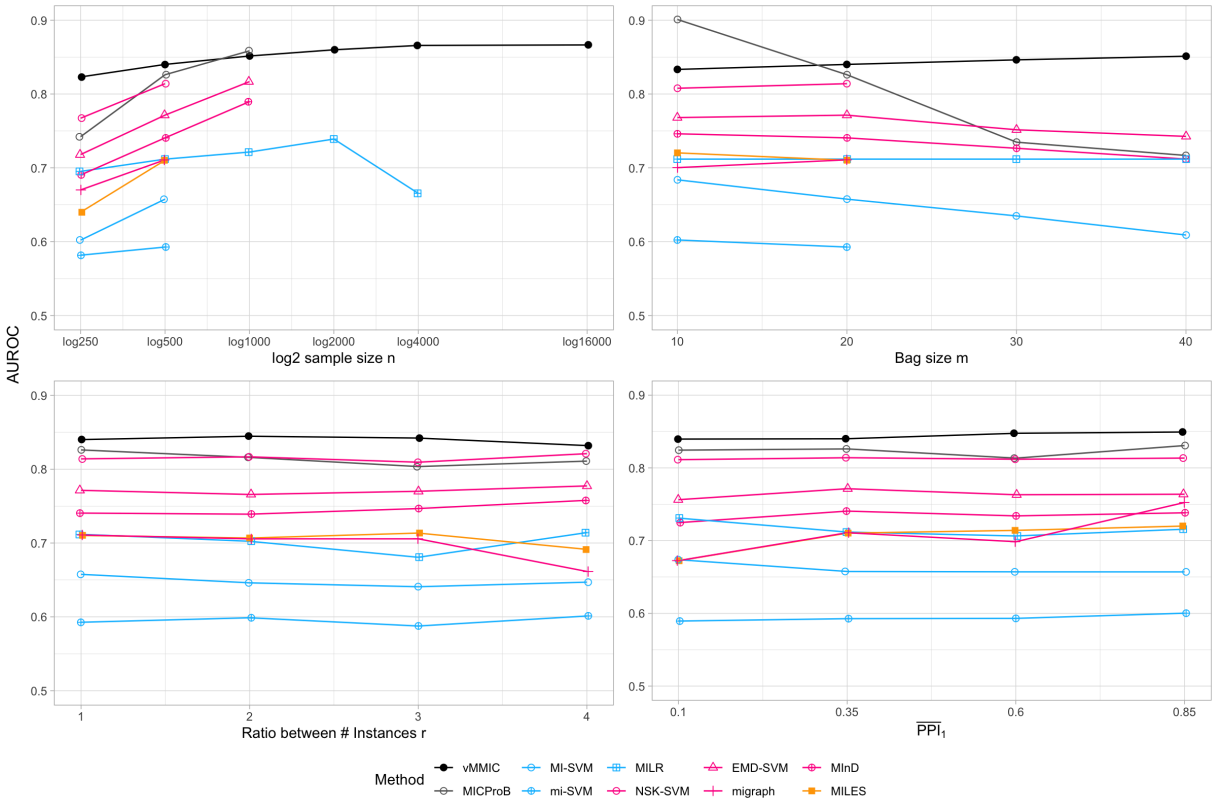
2.3.3. Results

The average AUROC values over 50 replicates of the 10 methods in various scenarios are displayed in Figure 2.2. vMMIC consistently demonstrates superior performance across most settings, followed by MICProB and NSK-SVM. EMD-SVM, MInD, and MILR are classified as middle performers, while the remaining methods exhibit less satisfactory performance.

As for factors that may influence the performance, we observe: (1) The performance of most methods is improved along with increasing sample size n as more information is available. (2) Different methods tend to exhibit diverse patterns as the bag size m increases. For instance, vMMIC shows an increasing trend with larger m, while MICProB and MI-SVM show the opposite trend. (3) The ratio between the numbers of instances from two modalities r and the proportion of primary instances in the second modality $\overline{\text{PPI}}_1$ seem to have minimal impact on performance, particularly for the top performers. As $\overline{\text{PPI}}_1$ increases, the average AUROC for most methods experiences slight improvements, which matches with our intuition as more signal is contained and captured.

We halted the execution of most methods on datasets with relatively large n and m due to excessive computational time. Some top performers such as MICProB and NSK-SVM are computationally intensive, and stop as early as $n = 1000$ or $n = 500$. Some middle performers such as MInD and MILR are more computationally efficient and continue until $n = 2000$ or $n = 4000$. In contrast, our vMMIC is the only method that is capable of handling datasets as extensive as $n = 16000$ bags within approximately 12 minutes. Figure 2.3 further displays how log2 average running time (in minutes) is impacted by sample size n and bag size m. Our novel approach, vMMIC, is the fastest while maintaining promising bag-level

Figure 2.2: Simulation evaluation: average AUROC for bag classification over 50 replicates using different MIC methods in various simulation scenarios. Benchmark methods are distinguished by color (black: vMMIC; grey: MICProB; blue: IS methods; pink: BS methods; orange: ES methods). A line that stops somewhere in middle indicates the running time on a single repetition of the associated method exceeds 1.5 hours at a specific setting (and beyond).



prediction. In the basic setting, vMMIC runs approximately 83 times faster than MICProB and NSK-SVM.

In Figure 2.4, we demonstrate the performance of vMMIC and MICProB in the identification of primary instances under various conditions. vMMIC consistently outperforms MICProB in determining which instances predominantly influence bag labels. Typically, vMMIC achieves an average AUROC exceeding 0.8 in primary instance identification, whereas MICProB's average AUROC fluctuates between 0.55 and 0.6. Notably, MICProB, originally designed for unimodal instances, experiences an ad-hoc adaptation to multimodal instances by appending or prefixing zeros, resulting in sub-optimal PIs identification. In contrast, vMMIC employs a formal model structure to identify PIs, thus exhibiting markedly improved performance. vMMIC's performance in instance classification remains largely unaffected by changes in sample size n and bag size m, while MICProB's effectiveness increases as n increases but decreases as m increases. Moreover, when the proportion of information from different modalities becomes more balanced (i.e., r decreases to 1), it becomes increasingly challenging for both vMMIC and MICProB to accurately identify primary instances from two informative modalities. In addition, when the $\overline{\text{PPI}}$ of one modality is fixed while that of another increases, the performance of both vMMIC and MICProB deteriorates, possibly due to increased imbalance between the two $\overline{\text{PPI}}$s.

2.4. Real-world Yelp Ratings

2.4.1. Description of Yelp ratings

The widespread use of social media platforms like Google, Yelp, and TripAdvisor has fundamentally altered consumer behavior, particularly in how people evaluate recreational facilities and products. These online reviews are influential for making recommendations, as they significantly influence consumer decisions [73]. The shift from text-centric posts

Figure 2.3: Simulation evaluation: the log2 average computational time (in minutes) under the setting of different sample size n and bag size m.

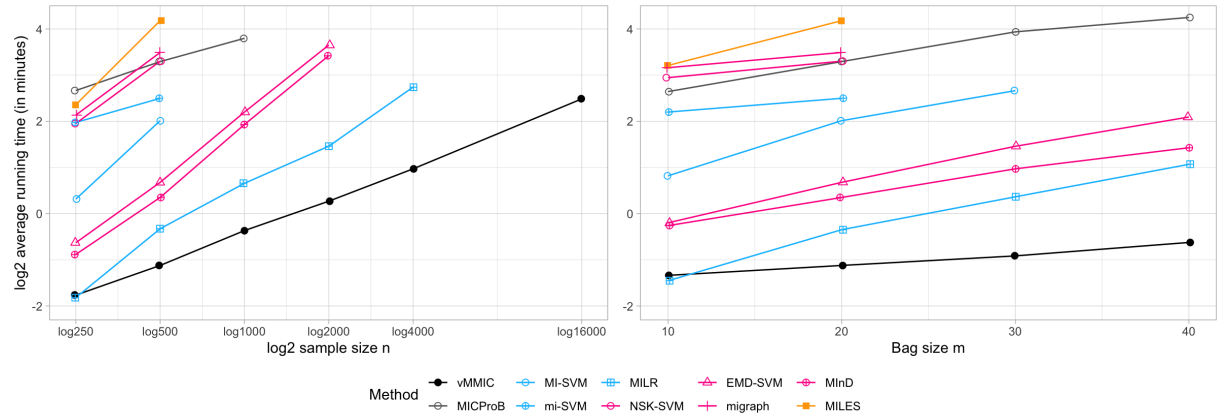
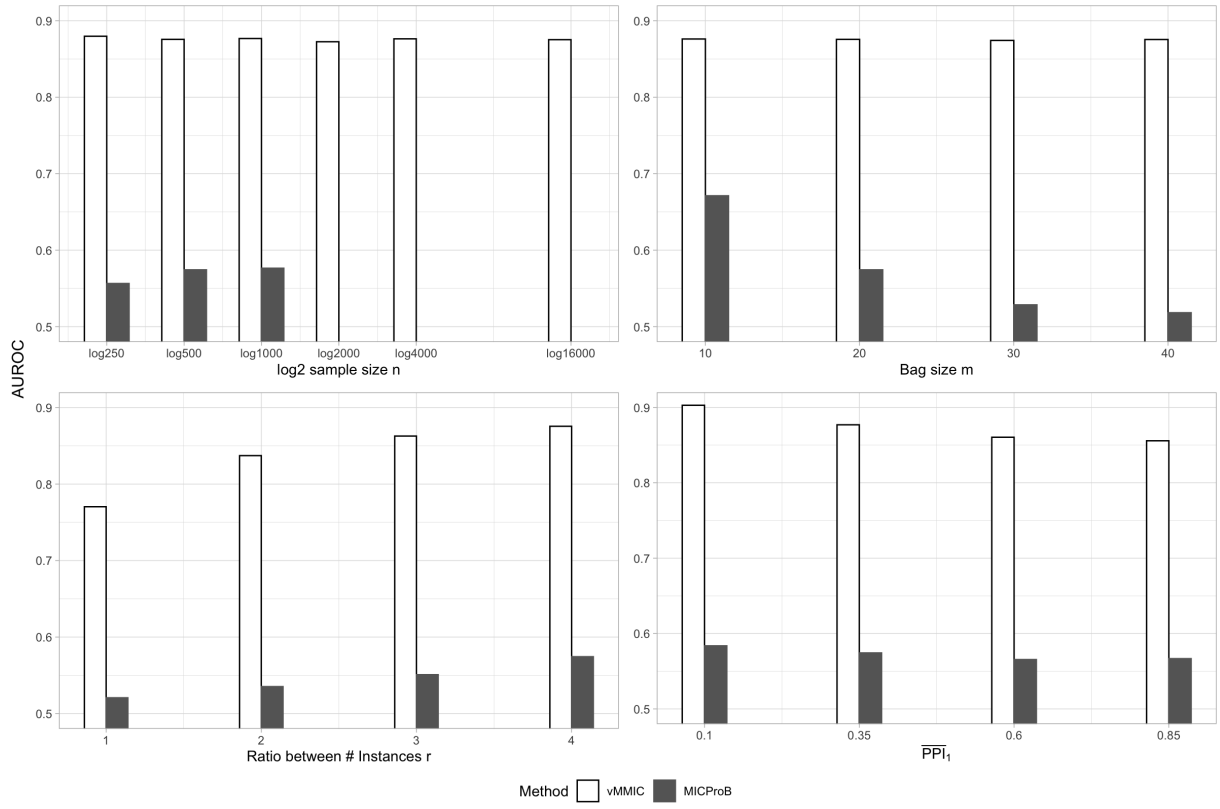


Figure 2.4: Simulation evaluation: the averaged AUROC over 50 replicates for identifying primary instances using vMMIC and MICProB in different scenarios. All other benchmark methods are not capable of identifying primary instances.



to visual-oriented posts arouses the need to analyze the effects of different components [74]. Review posts, along with user-provided images, both serve as rich sources of information [53]. In our study, we analyze the Yelp dataset, which includes 37,650 reviews of restaurants across five major U.S. cities: Boston, Chicago, Los Angeles, New York, and San Francisco [75]. Each review pairs textual content with at least three images and provides a rating on a scale of one to five stars. In line with our analytical focus on binary classification, we categorize the five-point rating scale into two distinct classes: positive for ratings of four or five stars, and negative for all other ratings. The ratio of positive to negative ratings stands at 2:3.

Sentences in reviews are viewed as instances from the textual modality, while images are treated as instances from a separate visual modality. Detailed descriptive statistics of the dataset are displayed in Table 2.1. On average, a review contains 11 sentences and three images, suggesting that text typically may carry more information than visuals. Both textual and image instances undergo preprocessing using the CLIP technique [76], which is acknowledged as one of the most state-of-art text and image embedding methods. The default configuration of CLIP projects sentences or images into a 512-dimensional space. To reduce the dimensionality, we implement a variational auto-encoder strategy [77] to transform text embeddings into 16 dimensions and image embeddings into 64 dimensions. Although a grid search for alternative combinations of dimension choices may yield marginally better performance, the overall impact on performance is not expected to be substantial. Similar to the bag aggregation trick applied in Section 2.3, we transform instances into 80-dimensional feature vectors by appending 64 zeros to textual instances and prefixing visual features with 16 zeros, thus adapting for unimodal MIC algorithms. To include computationally intensive MIC methods in the comparison study, as well as to examine the effects of information amount provided in the training set, we randomly select 500 ratings (300 negative ratings and 200 positive ratings) as the testing set, and try training sets with varying number of bags $n \in \{100, 200, 500, 1000, 2000, 5000, 20000, 37150\}$. As in Section 2.3, vMMIC and the other 9 benchmark MIC methods are included in the comparison. However, if a method's

running time exceeds 15 hours on a particular sample size n, it will be excluded from the analysis for that n and larger values of n.

2.4.2. Numerical comparison

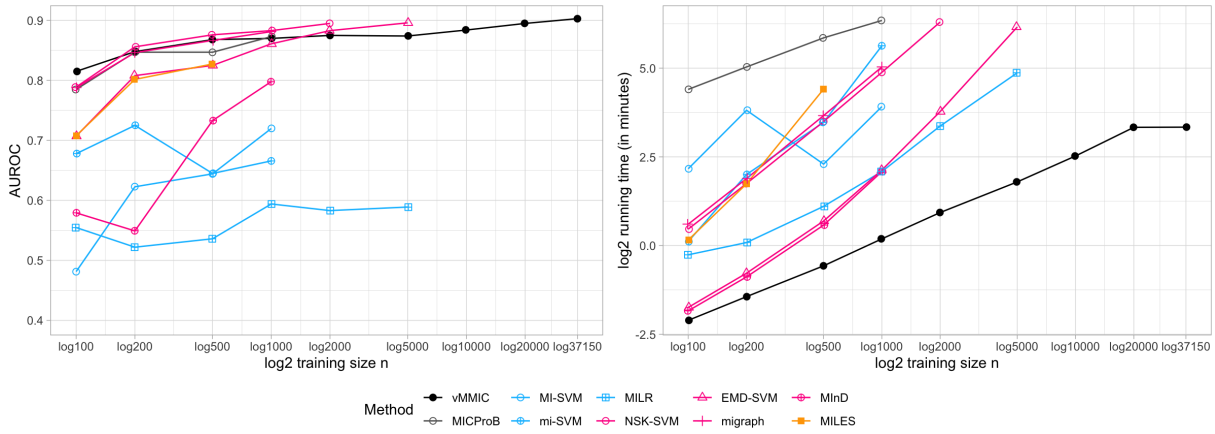
The results of AUROC and log2 running time (in minutes) against log2 training size n are shown in Figure 2.5. Consistent with the observations in Section 2.3, the majority of methods, particularly the top performers, exhibit improved performance with increasing sample size n. At a small sample size of 1000, vMMIC, MICProB, NSK-SVM, EMD-SVM, and miGraph are among the top performing methods. Notably, miGraph, which demonstrates moderate performance in simulated data, achieves similar results to NSK-SVM on Yelp data. MInD and MILES constitute the middle performing group, while the remaining methods show less satisfactory performance. MILR, despite performing well on simulated data, only achieves an AUROC between 0.5 and 0.6 across different values of n. Most methods stop early, with vMMIC, EMD-SVM, and MInD being exceptions. EMD-SVM and MInD continue until n reaches 5000. Particularly noteworthy is our vMMIC, the only method capable of handling datasets with as many as 10,000 bags, and even the entire set of 37,150 bags, in approximately 28 minutes.

2.4.3. Interpretability

Table 2.1: Summary statistics of textual and visual instance counts, and their ratio within each bag.

Statistic	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	Sd.
textual instance counts	1.00	6.00	11.00	13.52	18.00	104.00	10.61
visual instances counts	3.00	3.00	3.00	3.74	3.00	120.00	2.76
ratio between two counts	0.03	1.67	3.33	4.03	5.33	34.67	3.24

Figure 2.5: Yelp data: average AUROC and the log2 running time (in minutes) for predicting bag labels for each method across the log2 different sample size n.



As discussed in Section 2.1, only our vMMIC and MICProB [50] offer the functionality of identifying primary instances. However, as MICProB is designed for instances from a single modality, instances from multiple modalities are either appended or prefixed with zeros to create feature vectors of the same length. Consequently, the predictions of MICProB do not convey specific meanings, and therefore, we only present an example to demonstrate how vMMIC works in identifying instances from both modalities that contribute to a rating label. For this purpose, we select a rating with 2 stars for a Greek restaurant in Los Angeles. After preprocessing, the original rating comprises nine sentences and three images. In other words, for this particular bag, we have $\tilde{y} = 0$, $\tilde{m}^0 = 9$ and $\tilde{m}^1 = 3$. In Figure 2.6, sentences are ranked from the most relevant to the least relevant, so as images, based on the prediction of vMMIC. We can tell among all sentences, those with the strongest negative sentiment, such as so what went wrong? and combo meals are insanely pricey at around 19 dollars for gyro plate, side salad and greek potato, as well as the meal is about 7 dollars too much for me to consider going back, and for that type of price, I certainly expect more, are identified as most responsible for the negative bag label. Conversely, sentences without a discernible sentiment direction, such as the two sentences describing the layout in the restaurant and where the reviewer was originally from, are considered irrelevant to the bag label. Among the three images, the image depicting the gyro—the item that the reviewer found unsatisfactory—is

identified as the most relevant, while the other two images are viewed less relevant. This example demonstrates the superiority of vMMIC's ability to interpret its prediction, based on its transparent model structure.

2.5. Discussion

Multiple instance learning (MIL), especially for classification, has seen extensive research for its real-world value and reduced labeling costs. However, most existing MIL approaches are algorithm-driven and lack statistical grounding. Recognizing this, Xiong et al. [50] developed MICProB, a Bayesian MIC method that not only offers competitive performance and enhanced interpretability but also enables statistical inference and uncertainty quantification. While successful, MICProB and most existing methods focus on unimodal instances (featuring a single type of data) and become computationally intensive for large-scale data. Today, datasets often integrate information from multiple sources, as explored in Ma et al. [53]; Al-Tameemi et al. [54]; Sahasrabudhe et al. [58]; Xu et al. [55]; Tang et al. [56]. The extension of MIL into this multimodal domain remains under-explored, with the scalability of MIL algorithms facing even greater pressure due to the larger volume and complexity of multimodal data.

To address these gaps, we propose vMMIC, a novel multimodal MIC approach. vMMIC builds upon MICProB, modifying its two-level probit model for diverse instance sources. Thus, when handling unimodal data, the Bayesian model of vMMIC reduces to that of MICProB. However, vMMIC utilizes variational inference instead of MCMC for posterior computation, which achieves great scalability. To our knowledge, vMMIC represents the first statistical attempt in this area. As shown in Section 2.3 and Section 2.4, vMMIC demonstrates competitive performance and superior computational efficiency on both synthetic and real-world datasets. Moreover, its Bayesian foundation maintains the advantages of interpretability, statistical inference and uncertainty quantification that are often missing in algorithm-driven solutions.

Figure 2.6: Yelp data: A two-star rating for a Greek restaurant, referred to here as XXX, located in Los Angeles. Textual and visual elements are ranked from most to least influential in determining the rating. The rating also mentions another restaurant, referred to as YYY, for comparison.

- so what went wrong?
- combo meals are insanely pricey at around 19 dollars for gyro plate, side salad and greek potato.
- well, for the price, the gyro was just somewhat average.
- the meal is about 7 dollars too much for me to consider going back, and for that type of price, i certainly expect more.
- nothing beats the chicken and gyro at the athenian room in chicago and greek islands.
- the potatoes were overloaded with herbs and slightly chewy.
- i'd still prefer mediterranean gyro at just about any other place.
- XXX purports to be an authentic greek deli/restaurant and the layout reminds me of exactly what YYY does in glendale (but of course alternatively italian).
- coming from chicago, there is a huge greek community both in the city and the burbs (especially north and northwest), so when i heard about XXX knowing the market is somewhat sparse for authentic greek food, i immediately jumped to the idea.

most relevant image

intermediately relevant image

least relevant image

We point out several directions for future research. vMMIC relies on instances from multiple views to predict bag labels. In many real-world scenarios, in addition to instance-level information, bag-level information is also available and can be incorporated to improve the performance. For instance, to predict the positiveness of restaurant reviews, a user's overall sentiment in previous reviews, a restaurant's overall rating as well as the overall level of expenditure, etc., all can project useful information and can be included in the model with minor adaption. Moreover, instances from different sources are now treated independently. An interesting alternation is to consider the cross-modal interactions among instances within the same bag.

The rise of deep learning has sparked keen interest in utilizing neural networks for statistical inference. Our vMMIC model currently uses a two-level probit model: one level to identify bag labels and the second to pinpoint relevant instances. Combined with data augmentation and conditional conjugate priors, this structure enables efficient computation through closed-form solutions in our variational inference. However, probit models assume linear relationships via a probit link function, which can be restrictive for highly non-linear data. In such scenarios, replacing the probit model with a neural network for predicting primary instances could offer greater flexibility while maintaining interpretability (i.e., retaining the probit model for investigating how each feature influences bag labels). It would be intriguing to explore the use of variational autoencoders for efficient posterior computation in the context of MIL.

Appendix A

APPENDIX of CHAPTER 1

A.1. Derivation of Variational Distributions

In this section, we show the details about the derivation for each variational distribution.

A.1.1. Update $q(z_i)$

A.1.1.1. When $y_i = 1$

Assuming $y_i = 1$, then $p(y_i|z_i, \alpha_i) = (1-\alpha_i) \cdot 1(z_i > 0)$.

$$
\begin{aligned}
\ln q^*(z_i) &= \mathbb{E}_{q^*(-z_i)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\theta_i, a, b, \sigma^2, \alpha_i)}\left[\ln p\left(y_i \mid z_i, \alpha_i\right) + \ln p\left(z_i \mid \theta_i, a, b\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\alpha_i)}\left[\ln\left\{\left[(1-\alpha_i)\,1(z_i > 0)\right]^{y_i}\left[\alpha_i 1(z_i > 0) + 1(z_i \le 0)\right]^{1-y_i}\right\}\right] \\
&\quad + \mathbb{E}_{q^*(\theta_i, a, b)}\left[\ln\left\{\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(z_i - (a + b\theta_i)\right)^2\right\}\right\}\right] + \text{const} \\
&= \mathbb{E}_{q^*(\alpha_i)}\left[\ln(1-\alpha_i) \cdot 1(z_i > 0)\right] - \frac{1}{2}\mathbb{E}_{q^*(\theta_i, a, b)}\left[\left(z_i - (a + b\theta_i)\right)^2\right] + \text{const} \\
&= \mathbb{E}_{q^*(\alpha_i)}\left[\ln(1-\alpha_i) + \ln 1(z_i > 0)\right] - \frac{1}{2}\mathbb{E}_{q^*(\theta_i, a, b)}\left[z_i^2 - 2z_i(a + b\theta) + (a + b\theta_i)^2\right] + \text{const} \\
&= \ln 1(z_i > 0) - \frac{1}{2}z_i^2 + z_i\left(\mathbb{E}_{q^*(a)}[a] + \mathbb{E}_{q^*(b)}[b] \cdot \mathbb{E}_{q^*(\theta_i)}[\theta_i]\right) + \text{const}.
\end{aligned}
$$

Exponentiating this quantity and setting $m_i = \mathbb{E}_{q^*(a)}[a] + \mathbb{E}_{q^*(b)}[b] \cdot \mathbb{E}_{q^*(\theta_i)}[\theta_i]$, we obtain $q^*(z_i) \propto 1(z_i > 0) \cdot \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\}$. We observe that $q^*(z_i)$ follows a truncated normal

distribution with mean m_i and variance 1. Thus,

$$\mathbb{E}_{q^*(z_i)}[z_i] = m_i + \frac{\phi(m_i)}{1 - \Phi(-m_i)}.$$

A.1.1.2. When $y_i = 0$

Assuming $y_i = 0$, then $p(y_i \mid z_i, \alpha_i) = [\alpha_i 1(z_i > 0) + 1(z_i \leq 0)]^{1-y_i} = \alpha_i 1(z_i > 0) + 1(z_i \leq 0)$.

$$\begin{aligned}
\ln q^*(z_i) &= \mathbb{E}_{q^*(\alpha_i)}[\ln\{\alpha_i 1(z_i > 0) + 1(z_i \leq 0)\}] + \mathbb{E}_{q^*(\theta_i,a,b)}\left[\frac{1}{2}(z_i - (a + b\theta_i))^2\right] + \text{const} \\
&= \mathbb{E}_{q^*(\alpha_i)}[\ln\{\alpha_i 1(z_i > 0) + 1(z_i \leq 0)\}] - \frac{1}{2}z_i^2 + z_i \cdot m_i + \text{const.} \\
q^*(z_i) &\propto \exp\left\{\mathbb{E}_{q^*(\alpha_i)}[\ln\{\alpha_i 1(z_i > 0) + 1(z_i \leq 0)\}]\right\} \cdot \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\} \\
&\propto \begin{cases} \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\}, & \text{if } z_i \leq 0; \\ \exp\left\{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]\right\} \cdot \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\}, & \text{if } z_i > 0. \end{cases}
\end{aligned}$$

As we can tell from the pdf of $q^*(z_i)$, this is a "mutated normal" distribution by multiplying $\exp\left\{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]\right\}$ to the part on the right-hand side of the y-axis. As $\alpha_i \in (0, 1)$, $\exp\left\{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]\right\}$ is always smaller than 1. In other words, the part for positive z_i shrinks down while the part for the negative z_i is expanded.

Now we would like to find the normalizing factor C' which guarantees the integral of the "mutated normal" distribution $q^*(z_i)$ to be 1:

$$\begin{aligned}
C'\left(\int_{-\infty}^{0} \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\} dz_i + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]} \cdot \int_{0}^{\infty} \exp\left\{-\frac{1}{2}z_i^2 + z_i \cdot m_i\right\} dz_i\right) &= 1 \\
\sqrt{2\pi}e^{\frac{1}{2}m_i^2} \cdot C'\frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{0} \exp\left\{-\frac{1}{2}(z_i - m_i)^2\right\} dz_i + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]} \cdot \int_{0}^{\infty} \exp\left\{-\frac{1}{2}(z_i - m_i)^2\right\} dz_i\right) &= 1 \\
\sqrt{2\pi} \cdot C' \cdot e^{\frac{1}{2}m_i^2}\left(\Phi(-m_i) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}(1 - \Phi(-m_i))\right) &= 1.
\end{aligned}$$

Thus,

$$C' = \frac{e^{-\frac{1}{2}m_i^2}}{\sqrt{2\pi}\left[\Phi\left(-m_i\right) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}\left(1-\Phi\left(-m_i\right)\right)\right]}.$$

Therefore,

$$q^*\left(z_i\right) = \begin{cases} \dfrac{1}{\Phi\left(-m_i\right) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}\left(1-\Phi\left(-m_i\right)\right)} \cdot \dfrac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(z_i - m_i\right)^2\right\}, & \text{if } z_i \le 0; \\ \dfrac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{\Phi\left(-m_i\right) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}\left(1-\Phi\left(-m_i\right)\right)} \cdot \dfrac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(z_i - m_i\right)^2\right\}, & \text{if } z_i > 0. \end{cases}$$

The expectation of z_i when $y_i = 0$ is computed as following:

$$\begin{aligned} \mathbb{E}_{q^*(z_i)}\left[z_i\right] = & \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(z_i - m_i\right)^2\right\} \cdot z_i dz_i ① \\ & + \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(z_i - m_i\right)^2\right\} \cdot z_i dz_i ②. \end{aligned}$$

Then by substitution (define $x_i = z_i - m_i$), we have

$$\begin{aligned} ① &= \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}(z_i - m_i)^2\right\} \cdot z_i dz_i \\ &= \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_{-\infty}^{-m_i} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x_i^2\right\} \cdot (x_i + m_i) dx_i \\ &= \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-m_i} x_i \exp\left\{-\frac{1}{2}x_i^2\right\} dx_i + m_i \int_{-\infty}^{-m_i} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x_i^2\right\} dx_i\right] \\ &= \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[-\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x_i^2\right\}\Big|_{-\infty}^{-m_i} + m_i \int_{-\infty}^{-m_i} \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}x_i^2\right\} dx_i\right] \\ &= \frac{1}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[m_i \cdot t_i - \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}m_i^2\right\}\right], \end{aligned}$$

and

$$
\begin{aligned}
② &= \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_0^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(z_i - m_i\right)^2\right\} \cdot z_i dz_i \\
&= \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \int_{-m_i}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} \cdot \left(x_i + m_i\right) dx_i \\
&= \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[\frac{1}{\sqrt{2\pi}} \int_{-m_i}^{\infty} x_i \exp\left\{-\frac{1}{2}x_i^2\right\} dx_i + m_i \int_{-m_i}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} dx_i\right] \\
&= \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[-\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\}\Big|_{-m_i}^{\infty} + m_i \int_{-m_i}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x_i^2\right\} dx_i\right] \\
&= \frac{e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}} \left[m_i \cdot (1-t_i) + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}m_i^2\right\}\right].
\end{aligned}
$$

Combining ① and ②, we obtain

$$
\mathbb{E}_{q^*(z_i)}\left[z_i\right] = \frac{m_i \cdot t_i - \phi\left(m_i\right) + e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]} \cdot \left[m_i \cdot (1-t_i) + \phi\left(m_i\right)\right]}{t_i + (1-t_i)\, e^{\mathbb{E}_{q^*(\alpha_i)}[\ln\alpha_i]}}.
$$

A.1.2. Update $q^*\left(\sigma^2\right)$

$$
\begin{aligned}
\ln q^*\left(\sigma^2\right) &= \mathbb{E}_{q^*(-\sigma^2)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\ln \pi\left(\boldsymbol{\theta} \mid \sigma^2\right)\right] + \ln \pi\left(\sigma^2\right) + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\ln\left\{(2\pi)^{-n/2} \det\{\boldsymbol{U}\sigma^2\}^{-1/2} \exp\left\{-\frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\left(\boldsymbol{U}\sigma^2\right)^{-1}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)\right\}\right\}\right] \\
&\quad + \ln\left[\frac{\tau^{\tau}}{\Gamma(\tau)}\left(\sigma^2\right)^{-(\tau+1)} \cdot \exp\left\{-\frac{\tau}{\sigma^2}\right\}\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[-\frac{1}{2}\ln\left\{\det\left(\boldsymbol{U}\sigma^2\right)\right\} - \frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{\boldsymbol{U}^{-1}}{\sigma^2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right) - (\tau+1)\ln\sigma^2 - \frac{\tau}{\sigma^2}\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[-\frac{n}{2}\ln\sigma^2 - \frac{1}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{\boldsymbol{U}^{-1}}{\sigma^2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right) - (\tau+1)\ln\sigma^2 - \frac{\tau}{\sigma^2}\right] + \text{const} \\
&= -\frac{n}{2}\ln\sigma^2 - (\tau+1)\ln\sigma^2 - \frac{\tau}{\sigma^2} - \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{\boldsymbol{U}^{-1}}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)\right]\cdot\frac{1}{\sigma^2} + \text{const} \\
&= -\left(\frac{n}{2}+\tau+1\right)\ln\sigma^2 - \left\{\tau + \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{U^{-1}}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)\right]\right\}\cdot\frac{1}{\sigma^2} + \text{const},
\end{aligned}
$$

which is the log of an inverse gamma distribution. Exponentiate the term, we can get

$$
q^*\left(\sigma^2\right) \propto \frac{1}{\sigma^2}^{\frac{n}{2}+\tau+1} \cdot \exp\left\{-\frac{1}{\sigma^2}\left(\tau + \mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{U^{-1}}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)\right]\right)\right\}.
$$

We use g to denote $\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)^{\top}\frac{U^{-1}}{2}\left(\boldsymbol{\theta}-\boldsymbol{\theta_0}\right)\right]$. Therefore,

$$
q^*\left(\sigma^2\right) = IG\left(\sigma^2|\alpha_0, \beta_0\right), \text{ where } \alpha_0 = \tau + \frac{n}{2}, \beta_0 = \tau + g.
$$

Hence, $1/\sigma^2 \sim Gamma\left(\tau + \frac{n}{2}, \tau + g\right)$, then

$$
\mathbb{E}_{q^*(\sigma^2)}\left[\frac{1}{\sigma^2}\right] = \frac{\tau + \frac{n}{2}}{\tau + g}.
$$

A.1.3. Update $q^*(a)$

$$
\begin{aligned}
\ln q^*(a) &= \mathbb{E}_{q^*(-a)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},b)}\left[\ln p\left(\boldsymbol{z} \mid \boldsymbol{\theta}, a, b\right)\right] + \ln \pi\left(a\right) + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},b)}\left[\ln\left\{(2\pi)^{-n/2}\det\left(I_n\right)^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{z}-\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)\right)^T I_n\left(\boldsymbol{z}-\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)\right)\right\}\right\}\right] \\
&\quad + \ln\left\{\frac{1}{\sqrt{2\pi}\sigma_a}\exp\left\{-\frac{1}{2\sigma_a^2}a^2\right\}\right\} + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},b)}\left[-\frac{1}{2}\left(\boldsymbol{z}^T\boldsymbol{z} - 2\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)^T\boldsymbol{z} + \left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)^T\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)\right)\right] - \frac{1}{2\sigma_a^2}a^2 + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},b)}\left[-\frac{1}{2}\left(-2a\boldsymbol{z}^T\mathbf{1}_n + 2ab\boldsymbol{\theta}^T\mathbf{1}_n + na^2\right)\right] - \frac{1}{2\sigma_a^2}a^2 + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},b)}\left[a\boldsymbol{z}^T\mathbf{1}_n - ab\boldsymbol{\theta}^T\mathbf{1}_n - \frac{n}{2}a^2\right] - \frac{1}{2\sigma_a^2}a^2 + \text{const} \\
&= a\mathbb{E}_{q^*(\boldsymbol{z})}\left[\boldsymbol{z}\right]^T\mathbf{1}_n - a\mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbf{1}_n - \frac{n}{2}a^2 - \frac{1}{2\sigma_a^2}a^2 + \text{const} \\
&= -\frac{1}{2}\left(n+\frac{1}{\sigma_a^2}\right)a^2 + a\left(\mathbb{E}_{q^*(\boldsymbol{z})}\left[\boldsymbol{z}\right]^T\mathbf{1}_n - \mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbf{1}_n\right) + \text{const}.
\end{aligned}
$$

By completing the square term, we obtain

$$
\begin{aligned}
q^*\left(a\right) &= N\left(a \mid \mu_a, S_a\right), \\
\mu_a &= \frac{\sigma_a^2}{n\sigma_a^2+1}\cdot\left(\mathbb{E}_{q^*(\boldsymbol{z})}\left[\boldsymbol{z}\right]^T\mathbf{1}_n - \mathbb{E}_{q^*(b)}\left[b\right]\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbf{1}_n\right), \\
S_a &= \left(n+\frac{1}{\sigma_a^2}\right)^{-1} = \frac{\sigma_a^2}{n\sigma_a^2+1}.
\end{aligned}
$$

A.1.4. Update $q^*(b)$

$$
\begin{aligned}
\ln q^*(b) &= \mathbb{E}_{q^*(-b)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},a)}\left[\ln p(\boldsymbol{z}|\boldsymbol{\theta}, a, b)\right] + \ln\pi(b) + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},a)}\left[\ln\left\{(2\pi)^{-n/2}\det(I_n)^{-1/2}\exp\left\{-\frac{1}{2}\left(\boldsymbol{z}-(a\mathbf{1}_n+b\boldsymbol{\theta})\right)^T I_n\left(\boldsymbol{z}-(a\mathbf{1}_n+b\boldsymbol{\theta})\right)\right\}\right\}\right] \\
&\quad + \ln\left\{\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left\{-\frac{1}{2\sigma_b^2}b^2\right\}\right\} + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},a)}\left[-\frac{1}{2}\left(\boldsymbol{z}^T\boldsymbol{z} - 2\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)^T\boldsymbol{z} + \left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)^T\left(a\mathbf{1}_n+b\boldsymbol{\theta}\right)\right)\right] - \frac{1}{2\sigma_b^2}b^2 + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},a)}\left[-\frac{1}{2}\left(-2b\boldsymbol{z}^T\boldsymbol{\theta} + b^2\boldsymbol{\theta}^T\boldsymbol{\theta} + 2ab\boldsymbol{\theta}^T\mathbf{1}_n\right)\right] - \frac{1}{2\sigma_b^2}b^2 + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z},\boldsymbol{\theta},a)}\left[b\boldsymbol{\theta}^T\boldsymbol{z} - \frac{1}{2}b^2\boldsymbol{\theta}^T\boldsymbol{\theta} - ab\boldsymbol{\theta}^T\mathbf{1}_n\right] - \frac{1}{2\sigma_b^2}b^2 + \text{const} \\
&= -\frac{1}{2}\left(\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}^T\boldsymbol{\theta}\right] + \frac{1}{\sigma_b^2}\right)b^2 + \left(\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbb{E}_{q^*(\boldsymbol{z})}\left[\boldsymbol{z}\right] - \mathbb{E}_{q^*(a)}\left[a\right]\cdot\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbf{1}_n\right)b + \text{const},
\end{aligned}
$$

which is the log of an un-normalized normal distribution. By completing the square term, we obtain

$$
\begin{aligned}
q^*(b) &= N\left(b \mid \mu_b, S_b\right), \\
\mu_b &= S_b\cdot\left(\mathbb{E}_{q(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbb{E}_{q^*(\boldsymbol{z})}\left[\boldsymbol{z}\right] - \mathbb{E}_{q^*(a)}\left[a\right]\cdot\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}\right]^T\mathbf{1}_n\right), \\
S_b &= \left(\mathbb{E}_{q^*(\boldsymbol{\theta})}\left[\boldsymbol{\theta}^T\boldsymbol{\theta}\right] + \frac{1}{\sigma_b^2}\right)^{-1}, \text{where } \mathbb{E}_{q^*(\boldsymbol{\theta})}[\boldsymbol{\theta}^T\boldsymbol{\theta}] = tr(\boldsymbol{S}_{\boldsymbol{\theta}}) + \boldsymbol{\mu}_{\boldsymbol{\theta}}{}^T\boldsymbol{\mu}_{\boldsymbol{\theta}}.
\end{aligned}
$$

A.1.5. Update $q^*(\alpha_i)$

$$
\begin{aligned}
\ln q^*(\alpha_i) &= \mathbb{E}_{q^*(-\alpha_i)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^2, \boldsymbol{\alpha}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z})}\left[\ln p\left(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{\alpha}\right)\right] + \ln \pi\left(\boldsymbol{\alpha}\right) + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z})}\left[\ln\left\{\prod_{i=1}^{n} p\left(y_i \mid z_i, \alpha_i\right)\right\}\right] + \text{const} \\
&= \mathbb{E}_{q^*(\boldsymbol{z})}\left[\sum_{i=1}^{n} \ln p\left(y_i \mid z_i, \alpha_i\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(z_i)}\left[\ln\left\{\left[\left(1-\alpha_i\right) 1\left(z_i>0\right)\right]^{y_i}\left[\alpha_i 1\left(z_i>0\right)+1\left(z_i>0\right)\right]^{1-y_i}\right\}\right] + \text{const} \\
&= \mathbb{E}_{q^*(z_i)}\left[y_i \ln\left\{\left(1-\alpha_i\right) 1\left(z_i>0\right)\right\} + \left(1-y_i\right) \ln\left\{\alpha_i 1\left(z_i>0\right)+1\left(z_i>0\right)\right\}\right] + \text{const}.
\end{aligned}
$$

When $y_i = 1$, we have $1(z_i > 0) = 1$, then $\mathbb{E}_{q^*(z_i)}\left[y_i \ln\left\{\left(1-\alpha_i\right) 1\left(z_i>0\right)\right\}\right] = \mathbb{E}_{q^*(z_i)}\left[y_i \ln\left(1-\alpha_i\right)\right]$. When $y_i = 0$, if $z_i > 0$, $1(z_i > 0) = 1$, we have $\mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) \ln\left\{\alpha_i 1\left(z_i>0\right)+1\left(z_i>0\right)\right\}\right] = \mathbb{E}_{q^*(z_i)}\left[\ln \alpha_i\right]$; If $z_i \leq 0$, $1(z_i > 0) = 0$, then $\mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) \ln\left\{\alpha_i 1\left(z_i>0\right)+1\left(z_i>0\right)\right\}\right] = \mathbb{E}_{q^*(z_i)}\left[\ln\left(1\left(z_i \leq 0\right)\right)\right] = 0$. Therefore, regardless of the sign of z_i, when $y_i = 0$, we have

$$
\ln q^*\left(\alpha_i\right) = \mathbb{E}_{q^*(z_i)}\left[1\left(z_i>0\right) \ln \alpha_i\right] + \text{const} = \mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) 1\left(z_i>0\right) \ln \alpha_i\right] + \text{const}.
$$

Henceforth,

$$
\begin{aligned}
\ln q^*\left(\alpha_i\right) &= \mathbb{E}_{q^*(z_i)}\left[\left\{y_i \ln\left(1-\alpha_i\right) + \left(1-y_i\right) 1\left(z_i>0\right) \ln \alpha_i\right\}\right] + \text{const} \\
&= \left[\ln\left(1-\alpha_i\right)\right] \cdot y_i + \left[\ln \alpha_i\right] \cdot \mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) 1\left(z_i>0\right)\right] + \text{const}. \\
q^*\left(\alpha_i\right) &\propto \exp\left\{\left[\ln\left(1-\alpha_i\right)\right] y_i + \left[\ln \alpha_i\right] \cdot \mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) 1\left(z_i>0\right)\right]\right\} \\
&\propto \left(1-\alpha_i\right)^{y_i} \times \alpha_i^{\mathbb{E}_{q^*(z_i)}\left[\left(1-y_i\right) 1\left(z_i>0\right)\right]}.
\end{aligned}
$$

Obviously,

$$
\begin{aligned}
q^{*}\left(\alpha_{i}\right) &= Beta\left(\alpha_{i1}, \beta_{i1}\right), \\
\alpha_{i1} &= \mathbb{E}_{q^{*}\left(z_{i}\right)}\left[\left(1-y_{i}\right) 1\left(z_{i}>0\right)\right]+1, \\
\beta_{i1} &= y_{i}+1.
\end{aligned}
$$

Here when $y_i = 1$, $\mathbb{E}_{q^{*}(z_i)}\left[\left(1-y_{i}\right) 1\left(z_{i}>0\right)\right] = 0$; when $y_i = 0$, $\mathbb{E}_{q^{*}(z_i)}\left[\left(1-y_{i}\right) 1\left(z_{i}>0\right)\right] = \dfrac{e^{\mathbb{E}_{q^{*}(\alpha_i)}[\ln\alpha_i]}\left(1-t_{i}\right)}{t_{i}+e^{\mathbb{E}_{q^{*}(\alpha_i)}[\ln\alpha_i]}\left(1-t_{i}\right)}$, where $t_i = \Phi\left(-m_{i}\right)$.

A.2. Derivation of Variational Distributions in a Simplified Case $\alpha_i \equiv \alpha$

In the main text, the real-world data examples are computed assuming $\alpha_i = \alpha$ across $i = 1, \cdots, n$. With the simplification in model setup, the variational posteriors of parameters of θ_i, a, b, and σ^2 remain the same. As in the derivation for $q^{*}\left(z_{i}\right)$, we simply replace terms involving α_i with corresponding terms involving α. However, the derivation of $q^{*}\left(\alpha\right)$ is modified as outlined below.

$$
\begin{aligned}
\ln q^{*}(\alpha) &= \mathbb{E}_{q^{*}(-\alpha)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{z}, \boldsymbol{\theta}, a, b, \sigma^{2}, \alpha\right)\right]+\text{const} \\
&= \mathbb{E}_{q^{*}(\boldsymbol{z})}\left[\ln p\left(\boldsymbol{y} \mid \boldsymbol{z}, \alpha\right)\right]+\ln \pi\left(\alpha\right)+\text{const} \\
&= \mathbb{E}_{q^{*}(\boldsymbol{z})}\left[\ln\left\{\prod_{i=1}^{n} p\left(y_{i} \mid z_{i}, \alpha\right)\right\}\right]+\text{const} \\
&= \mathbb{E}_{q^{*}(\boldsymbol{z})}\left[\sum_{i=1}^{n} \ln p\left(y_{i} \mid z_{i}, \alpha\right)\right]+\text{const} \\
&= \mathbb{E}_{q^{*}(z_i)}\left[\sum_{i=1}^{n} \ln\left\{\left[\left(1-\alpha\right) 1\left(z_{i}>0\right)\right]^{y_{i}}\left[\alpha 1\left(z_{i}>0\right)+1\left(z_{i}>0\right)\right]^{1-y_{i}}\right\}\right]+\text{const} \\
&= \mathbb{E}_{q^{*}(z_i)}\left[\sum_{i=1}^{n} y_{i} \ln\left\{\left(1-\alpha\right) 1\left(z_{i}>0\right)\right\}+\left(1-y_{i}\right) \ln\left\{\alpha 1\left(z_{i}>0\right)+1\left(z_{i}>0\right)\right\}\right]+\text{const}.
\end{aligned}
$$

When $y_i = 1$, we have $1(z_i > 0) = 1$, then $\mathbb{E}_{q^{*}(z_i)}\left[y_{i} \ln\left\{\left(1-\alpha\right) 1\left(z_{i}>0\right)\right\}\right] = \mathbb{E}_{q^{*}(z_i)}\left[y_{i} \ln\left(1-\alpha\right)\right]$. When $y_i = 0$, we have $\mathbb{E}_{q^{*}(z_i)}\left[\left(1-y_{i}\right) \ln\left\{\alpha 1\left(z_{i}>0\right)+1\left(z_{i} \leq 0\right)\right\}\right]$ always equal to $\mathbb{E}_{q^{*}(z_i)}\left[\left(1-y_{i}\right) 1\left(z_{i}>0\right) \ln\alpha\right]$ regardless of the sign of z_i.

Henceforth,

$$\begin{aligned}
\ln q^{*}(\alpha) &= \mathbb{E}_{q^{*}(z_i)}\left[\sum_{i=1}^{n}\left\{y_i \ln(1-\alpha)+(1-y_i)\,1(z_i>0)\ln\alpha\right\}\right]+\text{const} \\
&= (\ln(1-\alpha))\cdot\sum_{i=1}^{n} y_i+(\ln\alpha)\cdot\sum_{i=1}^{n}\mathbb{E}_{q^{*}(z_i)}\left[(1-y_i)\,1(z_i>0)\right]+\text{const.} \\
q^{*}(\alpha_i) &\propto \exp\left\{(\ln(1-\alpha))\cdot\sum_{i=1}^{n} y_i+(\ln\alpha)\cdot\sum_{i=1}^{n}\mathbb{E}_{q^{*}(z_i)}\left[(1-y_i)\,1(z_i>0)\right]\right\} \\
&\propto (1-\alpha)^{\sum_{i=1}^{n} y_i}\times\alpha^{\sum_{i=1}^{n}\mathbb{E}_{q^{*}(z_i)}[(1-y_i)1(z_i>0)]}.
\end{aligned}$$

Obviously,

$$\begin{aligned}
q^{*}(\alpha) &= Beta(\alpha_1,\beta_1), \\
\alpha_1 &= \sum_{i=1}^{n}\mathbb{E}_{q^{*}(z_i)}\left[(1-y_i)\,1(z_i>0)\right]+1, \\
\beta_{i1} &= \sum_{i=1}^{n} y_i+1.
\end{aligned}$$

Here when $y_i=1$, $\mathbb{E}_{q^{*}(z_i)}\left[(1-y_i)\,1(z_i>0)\right]=0$; when $y_i=0$, $\mathbb{E}_{q^{*}(z_i)}\left[(1-y_i)\,1(z_i>0)\right]=\dfrac{e^{\mathbb{E}_{q^{*}(\alpha)}[\ln\alpha]}(1-t_i)}{t_i+e^{\mathbb{E}_{q^{*}(\alpha)}[\ln\alpha]}(1-t_i)}$.

A.3. Additional Real-world Data Results

A.3.1. Performance on PubMed articles

In Chapter 1, we include Table 1.4 which displays the precisions, recalls, and F-measures of various approaches for the PubMed articles. It is noteworthy that unsupervised approaches do not utilize partial label information, thus are not guaranteed to identify observed keywords. Therefore, here we force the observed keywords to be positives for those unsupervised methods and recalculate their metrics in Table 1.1.

Table 1.1: PubMed data: the comparison of precisions, recalls and F-measures of various keyword extraction approaches, by forcing the observed keywords to be positives. We report the results obtained with different FDR values γ.

FDR cutoff γ	**Precision**									
	VBSS	BSS	SS	TR	TpR	MR	PosR	TF-IDF	KPMiner	YAKE
0.05	**1**	*0.536*	0.975	0.969	0.932	0.938	0.963	0.969	0.969	0.938
0.1	**0.934**	*0.386*	0.875	0.865	0.828	0.813	0.865	0.880	0.870	0.833
0.15	**0.848**	*0.282*	0.82	0.802	0.756	0.756	0.797	0.816	0.806	0.774
0.2	**0.783**	*0.205*	0.757	0.753	0.684	0.680	0.725	0.773	0.773	0.704
0.25	0.675	*0.147*	0.666	0.659	0.581	0.565	0.604	0.675	**0.679**	0.610
0.3	**0.588**	*0.107*	0.591	0.578	0.494	0.468	0.495	0.578s	0.584	0.530
	Recall									
0.05	*0.329*	**0.616**	0.346	0.344	0.331	0.333	0.342	0.344	0.344	0.333
0.1	0.371	**0.702**	0.368	0.364	0.349	*0.342*	0.364	0.371	0.366	0.351
0.15	0.39	**0.763**	0.39	0.382	*0.360*	*0.360*	0.379	0.388	0.384	0.368
0.2	0.419	**0.816**	0.41	0.408	0.371	*0.368*	0.393	0.419	0.419	0.382
0.25	0.452	**0.871**	0.45	0.445	0.393	*0.382*	0.408	0.456	0.458	0.412
0.3	0.5	**0.91**	0.504	0.493	0.421	*0.399*	0.419	0.493	0.496	0.452
	F-measure									
0.05	0.495	**0.573**	0.511	0.508	*0.489*	0.492	0.505	0.508	0.508	0.492
0.1	**0.531**	0.498	0.519	0.512	0.491	*0.481*	0.512	0.522	0.515	0.494
0.15	**0.535**	*0.412*	0.529	0.517	0.487	0.487	0.514	0.526	0.521	0.499
0.2	**0.546**	*0.328*	0.532	0.529	0.481	0.478	0.509	0.543	0.543	0.495
0.25	0.541	*0.251*	0.537	0.531	0.469	0.455	0.487	0.545	**0.547**	0.492
0.3	0.541	*0.191*	**0.544**	0.533	0.454	0.431	0.454	0.533	0.537	0.488

A.3.2. Performance on Hulth

For the Hulth data, in addition to the time consumption comparison of VBSS and BSS included in the main text, we also compare their performance in Table 1.2. Here, the other existing methods are not included for comparison since Wang et al. [1] has shown that BSS was the winner among all for this dataset. VBSS consistently exhibits higher precision across various FDR control values γ, while BSS is characterized by superior recall. The overall performance, as assessed by the F-measure, gradually improves for VBSS, surpassing that of BSS as γ values increase.

Table 1.2: Hulth data: the comparison of precisions, recalls and F-measures of VBSS and BSS with different FDR values γ.

		FDR Control γ					
		0.05	0.1	0.15	0.2	0.25	0.3
Precision	VBSS	**0.971**	**0.886**	**0.774**	**0.657**	**0.548**	**0.447**
	BSS	0.801	0.693	0.596	0.493	0.394	0.362
Recall	VBSS	0.273	0.34	0.421	0.527	0.666	0.828
	BSS	**0.443**	**0.568**	**0.687**	**0.798**	**0.923**	**0.971**
F-Measures	VBSS	0.426	0.492	0.546	0.585	**0.601**	**0.581**
	BSS	**0.571**	**0.624**	**0.638**	**0.61**	0.552	0.528

Appendix B

APPENDIX of CHAPTER 2

B.1. Derivation of Variational Distributions

In this section, we show the details about the derivation for each variational distribution.

B.1.1. Update $q^*(\alpha)$

We define $l_i = (\delta_{i1}(1-w_{i1}), \cdots, \delta_{im_i}(1-w_{im_i}))^T$ and $l_i' = (\delta_{i1}w_{i1}, \cdots, \delta_{im_i}w_{im_i})^T$.

$$
\begin{aligned}
\ln q^*(\alpha) &= \mathbb{E}_{q^*(-\alpha)}\left[\ln p(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w})\right] + \text{const} \\
&= \mathbb{E}_{q^*(-\alpha)}\left[\sum_{i=1}^{n} \ln p(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i)\right] + \ln p\left(\alpha \mid \mu_\alpha, \sigma_\alpha^2\right) + \text{const} \\
&= \mathbb{E}_{q^*(-\alpha)}\left[\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(y_i^* - \left(\alpha + \sum_{j=1}^{m_i}\delta_{ij}(1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta} + \sum_{j=1}^{m_i}\delta_{ij}w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\right)\right)^2\right\}\right)\right] \\
&\quad + \ln\left(\frac{1}{\sqrt{2\pi\sigma_\alpha^2}}\exp\left\{-\frac{1}{2\sigma_\alpha^2}(\alpha-\mu_\alpha)^2\right\}\right) + \text{const} \\
&= -\frac{1}{2}\left(\frac{1}{\sigma_\alpha^2} + n\right)\alpha^2 \\
&\quad + \alpha\left(\frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i=1}^{n}\left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T\boldsymbol{X}_i\mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] - \mathbb{E}_{q^*(\boldsymbol{l}_i')}[\boldsymbol{l}_i']^T\boldsymbol{Z}_i\mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)\right) \\
&\quad + \text{const}.
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(\alpha) &= N(m_\alpha, V_\alpha), \\
V_\alpha &= \left(\frac{1}{\sigma_\alpha^2} + n\right)^{-1}, \\
m_\alpha &= V_\alpha \left(\frac{\mu_\alpha}{\sigma_\alpha^2} + \sum_{i=1}^{n} \left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] - \mathbb{E}_{q^*(\boldsymbol{l}_i')}[\boldsymbol{l}_i']^T \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)\right).
\end{aligned}
$$

B.1.2. Update $q^*(\beta)$

$$
\begin{aligned}
\ln q^*(\boldsymbol{\beta}) &= \mathbb{E}_{q^*(-\boldsymbol{\beta})}[\ln p(\boldsymbol{y}, \mathbf{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w})] + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\beta})}\left[\sum_{i=1}^{n} \ln p(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i)\right] + \ln p(\boldsymbol{\beta} \mid \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\beta})}\left[\sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[\left(y_i^* - \alpha - \sum_{j=1}^{m_i} \delta_{ij} w_{ij} \boldsymbol{z}_{ij}\boldsymbol{\gamma}\right) - \sum_{j=1}^{m_i} \delta_{ij}(1 - w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta}\right]^2\right\}\right)\right] \\
&\quad + \ln\left((2\pi)^{-\frac{d_0+1}{2}} \det(\Sigma_{\boldsymbol{\beta}})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^\top \Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\}\right) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\beta})}\left[\sum_{i=1}^{n} \left(y_i^* - \alpha - \boldsymbol{l}_i'^T \boldsymbol{Z}_i \boldsymbol{\gamma}\right) \cdot \left(\boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta}\right) - \frac{1}{2}\left(\boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta}\right) \cdot \left(\boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta}\right)\right] - \frac{1}{2}\boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\beta} + \boldsymbol{\beta}^\top \Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} \\
&\quad + \text{const} \\
&= -\frac{1}{2}\boldsymbol{\beta}^\top \left(\Sigma_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i^T\right] \boldsymbol{X}_i\right) \boldsymbol{\beta} \\
&\quad + \boldsymbol{\beta}^\top \left(\Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{n} \left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right) \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}[\boldsymbol{l}_i] - \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i'^T\right] \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right) \\
&\quad + \text{const}.
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(\boldsymbol{\beta}) &= N(\boldsymbol{m}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}), \\
\boldsymbol{V}_{\boldsymbol{\beta}} &= \left(\Sigma_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i^T\right] \boldsymbol{X}_i\right)^{-1}, \\
\boldsymbol{m}_{\boldsymbol{\beta}} &= \boldsymbol{V}_{\boldsymbol{\beta}} \left(\Sigma_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{n} \left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right) \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}[\boldsymbol{l}_i] - \boldsymbol{X}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i \boldsymbol{l}_i'^T\right] \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right).
\end{aligned}
$$

B.1.3. Update $q^*(\gamma)$

$$
\begin{aligned}
\ln q^*(\boldsymbol{\gamma}) &= \mathbb{E}_{q^*(-\boldsymbol{\gamma})}\left[\ln p(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w})\right] + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\gamma})}\left[\sum_{i=1}^{n} \ln p(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i)\right] + \ln p(\boldsymbol{\gamma} \mid \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\gamma})}\left[\sum_{i=1}^{n} \ln\ \left(\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\left(y_i^* - \alpha - \boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta}\right) - \boldsymbol{l}_i'^T \boldsymbol{Z}_i \boldsymbol{\gamma}\right)^2\right\}\right)\right] \\
&\quad + \ln\left((2\pi)^{-\frac{d_1+1}{2}} \det(\Sigma_{\boldsymbol{\gamma}})^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}})^\top \Sigma_{\boldsymbol{\gamma}}^{-1}(\boldsymbol{\gamma} - \boldsymbol{\mu}_{\boldsymbol{\gamma}})\right\}\right) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{\gamma})}\left[\sum_{i=1}^{n} \left(y_i^* - \alpha - \boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta}\right) \cdot \left(\boldsymbol{l}_i'^T \boldsymbol{Z}_i \boldsymbol{\gamma}\right) - \frac{1}{2}\left(\boldsymbol{l}_i'^T \boldsymbol{Z}_i \boldsymbol{\gamma}\right)^2\right] - \frac{1}{2}\boldsymbol{\gamma}^\top \Sigma_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\gamma} + \boldsymbol{\gamma}^\top \Sigma_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}} + \text{const} \\
&= -\frac{1}{2}\boldsymbol{\gamma}^\top \left(\Sigma_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^{n} \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i' \boldsymbol{l}_i'^T\right] \boldsymbol{Z}_i\right) \boldsymbol{\gamma} \\
&\quad + \boldsymbol{\gamma}^T \left(\Sigma_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{i=1}^{n} \left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right) \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i'\right] - \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i' \boldsymbol{l}_i^T\right] \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}]\right) \\
&\quad + \text{const}.
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(\boldsymbol{\gamma}) &= N(\boldsymbol{m}_{\boldsymbol{\gamma}}, \boldsymbol{V}_{\boldsymbol{\gamma}}), \\
\boldsymbol{V}_{\boldsymbol{\gamma}} &= \left(\Sigma_{\boldsymbol{\gamma}}^{-1} + \sum_{i=1}^{n} \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i' \boldsymbol{l}_i'^T\right] \boldsymbol{Z}_i\right)^{-1}, \\
\boldsymbol{m}_{\boldsymbol{\gamma}} &= \boldsymbol{V}_{\boldsymbol{\gamma}} \left(\Sigma_{\boldsymbol{\gamma}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\gamma}} + \sum_{i=1}^{n} \left(\mathbb{E}_{q^*(y_i^*)}[y_i^*] - \mathbb{E}_{q^*(\alpha)}[\alpha]\right) \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i'\right] - \boldsymbol{Z}_i^T \mathbb{E}_{q^*(\boldsymbol{\delta}_i)}\left[\boldsymbol{l}_i' \boldsymbol{l}_i^T\right] \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}]\right).
\end{aligned}
$$

B.1.4. Update $q^*(a)$

$$
\begin{aligned}
\ln q^*(a) &= \mathbb{E}_{q^*(-a)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(-a)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln p\left(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right)\right] + \ln p\left(a \mid \mu_a, \sigma_a^2\right) + \text{const} \\
&= \mathbb{E}_{q^*(-a)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[U_{ij} - \left(\left(1-w_{ij}\right)\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right]^2\right\}\right)\right] \\
&\quad + \ln\left(\frac{1}{\sqrt{2\pi\sigma_a^2}}\exp\left\{-\frac{1}{2\sigma_a^2}\left(a - \mu_a\right)^2\right\}\right) + \text{const} \\
&= -\frac{1}{2}\left(\frac{1}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1 - w_{ij}\right)\right)a^2 \\
&\quad + a\left(\frac{\mu_a}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1 - w_{ij}\right)\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \boldsymbol{x}_{ij}\mathbb{E}_{q^*(\boldsymbol{b})}\left[\boldsymbol{b}\right]\right)\right) + \text{const.}
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(a) &= N\left(m_a, V_a\right), \\
V_a &= \left(\frac{1}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1 - w_{ij}\right)\right)^{-1}, \\
m_a &= V_a\left(\frac{\mu_a}{\sigma_a^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1 - w_{ij}\right)\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \boldsymbol{x}_{ij}\mathbb{E}_{q^*(\boldsymbol{b})}\left[\boldsymbol{b}\right]\right)\right).
\end{aligned}
$$

B.1.5. Update $q^*(b)$

$$
\begin{aligned}
\ln q^*(\boldsymbol{b}) &= \mathbb{E}_{q^*(-\boldsymbol{b})}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{b})}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln p\left(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right)\right] + \ln p\left(\boldsymbol{b} \mid \boldsymbol{\mu}_{\boldsymbol{b}}, \Sigma_{\boldsymbol{b}}\right) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{b})}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[U_{ij} - \left(\left(1-w_{ij}\right)\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right]^2\right\}\right)\right] \\
&\quad + \ln\left((2\pi)^{-\frac{d_0+1}{2}}\det\left(\Sigma_{\boldsymbol{b}}\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{b}-\boldsymbol{\mu}_{\boldsymbol{b}}\right)^T\Sigma_{\boldsymbol{b}}^{-1}\left(\boldsymbol{b}-\boldsymbol{\mu}_{\boldsymbol{b}}\right)\right\}\right) + \text{const} \\
&= -\frac{1}{2}\boldsymbol{b}^T\left(\Sigma_{\boldsymbol{b}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1-w_{ij}\right)\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ij}\right)\boldsymbol{b} \\
&\quad + \boldsymbol{b}^T\left(\Sigma_{\boldsymbol{b}}^{-1}\boldsymbol{\mu}_{\boldsymbol{b}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}^T\left(1-w_{ij}\right)\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \mathbb{E}_{q^*(a)}\left[a\right]\right)\right) + \text{const.}
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(\boldsymbol{b}) &= N\left(\boldsymbol{m}_{\boldsymbol{b}}, \boldsymbol{V}_{\boldsymbol{b}}\right), \\
\boldsymbol{V}_{\boldsymbol{b}} &= \left(\Sigma_{\boldsymbol{b}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\left(1-w_{ij}\right)\boldsymbol{x}_{ij}^T\boldsymbol{x}_{ij}\right)^{-1}, \\
\boldsymbol{m}_{\boldsymbol{b}} &= \boldsymbol{V}_{\boldsymbol{b}}\left(\Sigma_{\boldsymbol{b}}^{-1}\boldsymbol{\mu}_{\boldsymbol{b}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{x}_{ij}^T\left(1-w_{ij}\right)\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \mathbb{E}_{q^*(a)}\left[a\right]\right)\right).
\end{aligned}
$$

B.1.6. Update $q^*(c)$

$$
\begin{aligned}
\ln q^*(c) &= \mathbb{E}_{q^*(-c)}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(-c)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\, p\left(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right)\right] + \ln\, p\left(c \mid \mu_c, \sigma_c^2\right) + \text{const} \\
&= \mathbb{E}_{q^*(-c)}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(U_{ij} - \left(\left(1-w_{ij}\right)\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^2\right\}\right)\right] \\
&\quad + \ln\left(\frac{1}{\sqrt{2\pi\sigma_c^2}}\exp\left\{-\frac{1}{2\sigma_c^2}\left(c - \mu_c\right)^2\right\}\right) + \text{const} \\
&= -\frac{1}{2}\left(\frac{1}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\right)c^2 \\
&\quad + c\left(\frac{\mu_c}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\cdot\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \boldsymbol{z}_{ij}\mathbb{E}_{q^*(\boldsymbol{d})}\left[\boldsymbol{d}\right]\right)\right) + \text{const.}
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(c) &= N\left(m_c, V_c\right), \\
V_c &= \left(\frac{1}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\right)^{-1}, \\
m_c &= V_c\left(\frac{\mu_c}{\sigma_c^2} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\cdot\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \boldsymbol{z}_{ij}\mathbb{E}_{q^*(\boldsymbol{d})}\left[\boldsymbol{d}\right]\right)\right).
\end{aligned}
$$

B.1.7. Update $q^*(d)$

$$
\begin{aligned}
\ln q^*(\boldsymbol{d}) &= \mathbb{E}_{q^*(-\boldsymbol{d})}\left[\ln p\left(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w}\right)\right] + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{d})}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln p\left(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}\right)\right] + \ln p\left(\boldsymbol{d} \mid \boldsymbol{\mu}_{\boldsymbol{d}}, \Sigma_{\boldsymbol{d}}\right) + \text{const} \\
&= \mathbb{E}_{q^*(-\boldsymbol{d})}\left[\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(U_{ij} - \left(\left(1-w_{ij}\right)\left(a + \boldsymbol{x}_{ij}\boldsymbol{b}\right) + w_{ij}\left(c + \boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^2\right\}\right)\right] \\
&\quad + \ln\left((2\pi)^{-\frac{d_1+1}{2}}\det\left(\Sigma_{\boldsymbol{d}}\right)^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{d}-\boldsymbol{\mu}_{\boldsymbol{d}}\right)^\top \Sigma_{\boldsymbol{d}}^{-1}\left(\boldsymbol{d}-\boldsymbol{\mu}_{\boldsymbol{d}}\right)\right\}\right) + \text{const} \\
&= -\frac{1}{2}\boldsymbol{d}^\top\left(\Sigma_{\boldsymbol{d}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\boldsymbol{z}_{ij}^\top\boldsymbol{z}_{ij}\right)\boldsymbol{d} \\
&\quad + \boldsymbol{d}^\top\left(\Sigma_{\boldsymbol{d}}^{-1}\boldsymbol{\mu}_{\boldsymbol{d}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{z}_{ij}^\top w_{ij}\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \mathbb{E}_{q^*(c)}\left[c\right]\right)\right) + \text{const.}
\end{aligned}
$$

Therefore, we conclude

$$
\begin{aligned}
q^*(\boldsymbol{d}) &= N\left(\boldsymbol{m}_{\boldsymbol{d}}, \boldsymbol{V}_{\boldsymbol{d}}\right), \\
\boldsymbol{V}_{\boldsymbol{d}} &= \left(\Sigma_{\boldsymbol{d}}^{-1} + \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}\boldsymbol{z}_{ij}^\top\boldsymbol{z}_{ij}\right)^{-1}, \\
\boldsymbol{m}_{\boldsymbol{d}} &= \boldsymbol{V}_{\boldsymbol{d}}\left(\Sigma_{\boldsymbol{d}}^{-1}\boldsymbol{\mu}_{\boldsymbol{d}} + \sum_{i=1}^{n}\sum_{j=1}^{m_i}\boldsymbol{z}_{ij}^\top w_{ij}\left(\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] - \mathbb{E}_{q^*(c)}\left[c\right]\right)\right).
\end{aligned}
$$

B.1.8. Update $q^*(y_i^*)$

$$
\begin{aligned}
\ln q^*(y_i^*) &= \mathbb{E}_{q^*(-y_i^*)}\left[\ln p(\boldsymbol{y}, \boldsymbol{y}^*, \boldsymbol{\Delta}, \alpha, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{U}, a, \boldsymbol{b}, c, \boldsymbol{d} \mid \boldsymbol{X}, \boldsymbol{Z}, \boldsymbol{w})\right] + \text{const} \\
&= \mathbb{E}_{q^*(-y_i^*)}\left[\ln p(y_i \mid y_i^*) + \ln\, p(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i)\right] + \text{const} \\
&= \mathbb{E}_{q^*(-y_i^*)}\left[\ln\, 1(y_i^* > 0)^{\,y_i} 1(y_i^* \le 0)^{\,1-y_i}\right] \\
&\quad + \mathbb{E}_{q^*(-y_i^*)}\left[\ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(y_i^* - \left(\alpha + \boldsymbol{l}_i^T \boldsymbol{X}_i \boldsymbol{\beta} + \boldsymbol{l}_i'^T \boldsymbol{Z}_i \boldsymbol{\gamma}\right)\right)^2\right\}\right)\right] + \text{const} \\
&= y_i \ln\, 1(y_i^* > 0) + (1 - y_i) \ln\, 1(y_i^* \le 0) \\
&\quad - \frac{1}{2} y_i^{*2} + y_i^* \left(\mathbb{E}_{q^*(\alpha)}[\alpha] + \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] + \mathbb{E}_{q^*(\boldsymbol{l}_i')}\left[\boldsymbol{l}_i'\right]^T \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right) + \text{const}.
\end{aligned}
$$

$$
q^*(y_i^*) \propto \begin{cases} 1(y_i^* > 0) \cdot \exp\left\{-\frac{1}{2} y_i^{*2} + y_i^* \left(\mathbb{E}_{q^*(\alpha)}[\alpha] + \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] + \mathbb{E}_{q^*(\boldsymbol{l}_i')}[\boldsymbol{l}_i']^T \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)\right\}, & \text{if } y_i = 1; \\ 1(y_i^* \le 0) \cdot \exp\left\{-\frac{1}{2} y_i^{*2} + y_i^* \left(\mathbb{E}_{q^*(\alpha)}[\alpha] + \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]^T \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] + \mathbb{E}_{q^*(\boldsymbol{l}_i')}[\boldsymbol{l}_i']^T \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]\right)\right\}, & \text{if } y_i = 0. \end{cases}
$$

Therefore, we conclude

$$
q^*(y_i^*) = \begin{cases} TN_+\left(m_{y_i^*}, 1\right), & \text{if } y_i = 1, \\ TN_-\left(m_{y_i^*}, 1\right), & \text{if } y_i = 0, \end{cases}
$$

where $m_{y_i^*} = \mathbb{E}_{q^*(\alpha)}[\alpha] + \mathbb{E}_{q^*(\boldsymbol{l}_i)}[\boldsymbol{l}_i]\, \boldsymbol{X}_i \mathbb{E}_{q^*(\boldsymbol{\beta})}[\boldsymbol{\beta}] + \mathbb{E}_{q^*(\boldsymbol{l}_i')}[\boldsymbol{l}_i']\, \boldsymbol{Z}_i \mathbb{E}_{q^*(\boldsymbol{\gamma})}[\boldsymbol{\gamma}]$.

We have $\mathbb{E}_{q\left(y_i^*\right)}[y_i^*] = \begin{cases} m_{y_i^*} + \phi(-m_{y_i^*}) / \left[1 - \Phi\left(-m_{y_i^*}\right)\right], & \text{if } y_i = 1; \\ m_{y_i^*} - \phi(-m_{y_i^*}) / \Phi\left(-m_{y_i^*}\right), & \text{if } y_i = 0. \end{cases}$

B.1.9. Update $q^*(U_{ij})$

It is straightforward to obtain

$$
q(U_{ij} \mid \delta_{ij}) \propto \begin{cases} 1(\delta_{ij} = 1) \cdot 1(U_{ij} > 0) \cdot p(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}), & \text{if } \delta_{ij} = 1; \\ 1(\delta_{ij} = 0) \cdot 1(U_{ij} \le 0) \cdot p(U_{ij} \mid a, \boldsymbol{b}, c, \boldsymbol{d}, \boldsymbol{x}_{ij}, \boldsymbol{z}_{ij}, w_{ij}), & \text{if } \delta_{ij} = 0. \end{cases}
$$

In the next subsection, we know δ_{ij} follows a Bernoulli distribution with mean $\frac{A}{A+B}$. Hence,

$$q\left(U_{ij}\right)=\begin{cases}\frac{A}{A+B}\cdot\frac{1}{\sqrt{2\pi}}\exp\ \left\{-\frac{1}{2}\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right\}, & \text{if } U_{ij}>0;\\ \frac{B}{A+B}\cdot\frac{1}{\sqrt{2\pi}}\exp\ \left\{-\frac{1}{2}\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right\}, & \text{if } U_{ij}\leq 0.\end{cases}$$

It turns out that, $q\left(U_{ij}\right)$ is "merged" by two truncated normals. When $U_{ij}>0$, the normalizing constant is $\frac{A}{A+B}$, and when $U_{ij}\leq 0$, the normalizing constant is $\frac{B}{A+B}$.

To find the optimal variational posterior of U_{ij}, we take the expectations of $\ln q\left(U_{ij}\right)$ over all parameters except for U_{ij}.

- If $U_{ij}>0$,

$$\begin{aligned}\ln q^{*}\left(U_{ij}\right)&=\mathbb{E}_{q^{*}(-U_{ij})}\left[\ln\ \left(\frac{A}{A+B}\cdot\frac{1}{\sqrt{2\pi}}\exp\ \left\{-\frac{1}{2}\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right\}\right)\right]\\&=\ln\ \frac{A}{A+B}-\ln\ \sqrt{2\pi}-\frac{1}{2}\mathbb{E}_{q^{*}(-U_{ij})}\left[\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right]\\&=\ln\ \frac{A}{A+B}-\ln\ \sqrt{2\pi}-\frac{1}{2}\left\{U_{ij}^{2}-2U_{ij}\cdot m_{U_{ij}}+C'\right\},\end{aligned}$$

  where $m_{U_{ij}}=\left(1-w_{ij}\right)\left(\mathbb{E}_{q^{*}(a)}\left[a\right]+\boldsymbol{x}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{b})}\left[\boldsymbol{b}\right]\right)+w_{ij}\left(\mathbb{E}_{q^{*}(c)}\left[c\right]+\boldsymbol{z}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{d})}\left[\boldsymbol{d}\right]\right)$, and

$$\begin{aligned}C'=&-\frac{1}{2}\left\{\left(1-w_{ij}\right)\left(\mathbb{E}_{q^{*}(a)}\left[a^{2}\right]+2\mathbb{E}_{q^{*}(a)}\left[a\right]\cdot\boldsymbol{x}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{b})}\left[\boldsymbol{b}\right]+\boldsymbol{x}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{b})}\left[\boldsymbol{b}\boldsymbol{b}^{T}\right]\boldsymbol{x}_{ij}\right)\right\}\\&+w_{ij}\left(1-w_{ij}\right)\left(\mathbb{E}_{q^{*}(a)}\left[a\right]+\boldsymbol{x}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{b})}\left[\boldsymbol{b}\right]\right)\left(\mathbb{E}_{q^{*}(c)}\left[c\right]+\boldsymbol{z}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{d})}\left[\boldsymbol{d}\right]\right)\\&-\frac{1}{2}w_{ij}\left(\mathbb{E}_{q^{*}(c)}\left[c^{2}\right]+2\mathbb{E}_{q^{*}(c)}\left[c\right]\cdot\boldsymbol{z}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{d})}\left[\boldsymbol{d}\right]+\boldsymbol{z}_{ij}\mathbb{E}_{q^{*}(\boldsymbol{d})}\left[\boldsymbol{d}\boldsymbol{d}^{T}\right]\boldsymbol{z}_{ij}\right).\end{aligned}$$

- If $U_{ij}\leq 0$,

$$\begin{aligned}\ln q^{*}\left(U_{ij}\right)&=\mathbb{E}_{q^{*}(-U_{ij})}\left[\ln\ \left(\frac{B}{A+B}\cdot\frac{1}{\sqrt{2\pi}}\exp\ \left\{-\frac{1}{2}\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right\}\right)\right]\\&=\ln\ \frac{B}{A+B}-\ln\ \sqrt{2\pi}-\frac{1}{2}\mathbb{E}_{q^{*}(-U_{ij})}\left[\left(U_{ij}-\left(\left(1-w_{ij}\right)\left(a+\boldsymbol{x}_{ij}\boldsymbol{b}\right)+w_{ij}\left(c+\boldsymbol{z}_{ij}\boldsymbol{d}\right)\right)\right)^{2}\right]\\&=\ln\ \frac{B}{A+B}-\ln\ \sqrt{2\pi}-\frac{1}{2}\left\{U_{ij}^{2}-2U_{ij}\cdot m_{U_{ij}}+C'\right\}.\end{aligned}$$

It is equivalent to

$$q^*\left(U_{ij}\right) \propto \begin{cases} \frac{A}{A+B} \cdot \frac{1}{\sqrt{2\pi}} \exp\ \left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\}, & \text{if } U_{ij}>0; \\ \frac{B}{A+B} \cdot \frac{1}{\sqrt{2\pi}} \exp\ \left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\}, & \text{if } U_{ij} \leq 0. \end{cases}$$

We aim to find a universal normalizing factor C^0such that

$$C^0 \cdot \left(\frac{B}{A+B} \cdot \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij} + \frac{A}{A+B} \cdot \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij}\right) = 1.$$

It follows by

$$\begin{aligned} C^0 \cdot \left(\frac{B}{A+B} \cdot \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij} + \frac{A}{A+B} \cdot \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij}\right) &= 1 \\ C^0 \cdot \left(\frac{B}{A+B} \cdot \Phi\left(-m_{U_{ij}}\right) + \frac{A}{A+B} \cdot \left(1-\Phi\left(-m_{U_{ij}}\right)\right)\right) &= 1, \end{aligned}$$

so we get $C^0 = \frac{A+B}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)}$.

Therefore,

$$q^*\left(U_{ij}\right) = \begin{cases} \dfrac{A}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot \dfrac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\}, & \text{if } U_{ij}>0; \\ \dfrac{B}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot \dfrac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\}, & \text{if } U_{ij} \leq 0. \end{cases}$$

The expectation of U_{ij} is computed as following:

$$\begin{aligned} \mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] = & \frac{B}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} \cdot U_{ij} dU_{ij} ① \\ & + \frac{A}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \int_{0}^{+\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} \cdot U_{ij} dU_{ij} ②. \end{aligned}$$

Then by substitution (define $V_{ij} = U_{ij} - m_{U_{ij}}$),

$$\begin{aligned}
① &= \frac{B}{A + (B - A) \cdot \Phi\left(-m_{U_{ij}}\right)} \int_{-\infty}^{0} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij} - m_{U_{ij}}\right)^2\right\} \cdot U_{ij} dU_{ij} \\
&= \frac{B}{A + (B - A) \cdot \Phi\left(-m_{U_{ij}}\right)} \left[m_{U_{ij}} \cdot \Phi\left(-m_{U_{ij}}\right) - \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} m_{U_{ij}}^2\right\}\right],
\end{aligned}$$

and

$$\begin{aligned}
② &= \frac{A}{A + (B - A) \cdot \Phi\left(-m_{U_{ij}}\right)} \int_{0}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij} - m_{U_{ij}}\right)^2\right\} \cdot U_{ij} dU_{ij} \\
&= \frac{A}{A + (B - A) \cdot \Phi\left(-m_{U_{ij}}\right)} \left[m_{U_{ij}} \cdot \left(1 - \Phi\left(-m_{U_{ij}}\right)\right) + \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} m_{U_{ij}}^2\right\}\right].
\end{aligned}$$

Combine ① and ②, we obtain

$$\mathbb{E}_{q^*(U_{ij})}\left[U_{ij}\right] = \frac{(B - A) \cdot m_{U_{ij}} \cdot \Phi\left(-m_{U_{ij}}\right) + A \cdot m_{U_{ij}} + \frac{A-B}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2} m_{U_{ij}}^2\right\}}{A + (B - A) \cdot \Phi\left(-m_{U_{ij}}\right)}.$$

B.1.10. Update $q^*\left(\delta_{ij}\right)$

It is straightforward to obtain

$$q\left(\delta_{ij} \mid U_{ij}\right) \propto \begin{cases} 1\left(\delta_{ij} = 1\right) \cdot 1\left(U_{ij} > 0\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right), & \text{if } U_{ij} > 0; \\ 1\left(\delta_{ij} = 0\right) \cdot 1\left(U_{ij} \leq 0\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right), & \text{if } U_{ij} \leq 0. \end{cases}$$

- If $\delta_{ij} = 1$,

$$
\begin{aligned}
q\left(\delta_{ij}\right) &\propto \int_{-\infty}^{+\infty} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} \\
&\propto \int_{-\infty}^{0} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} + \int_{0}^{+\infty} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} \\
&\propto \int_{0}^{+\infty} 1\left(\delta_{ij}=1\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \cdot q\left(U_{ij}\right) dU_{ij} \\
&\propto 1\left(\delta_{ij}=1\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \\
&\quad \cdot \int_{0}^{+\infty} \frac{A}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij} \\
&\propto \frac{A\left(1-\Phi\left(-m_{U_{ij}}\right)\right)}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot 1\left(\delta_{ij}=1\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \\
&\propto \frac{A\left(1-\Phi\left(-m_{U_{ij}}\right)\right)}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot \delta_{ij} \cdot \frac{1}{\sqrt{2\pi}} \\
&\quad \cdot \exp\left\{-\frac{1}{2}\left[y_i^* - \left(\alpha + \delta_{ij}(1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta} + \delta_{ij}w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma} + \sum_{j'\neq j}^{m_i} \delta_{ij'}(1-w_{ij'})\boldsymbol{x}_{ij'}\boldsymbol{\beta} + \sum_{j'\neq j}^{m_i} \delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\right]^2\right\}.
\end{aligned}
$$

- When $\delta_{ij} = 0$,

$$
\begin{aligned}
q\left(\delta_{ij}\right) &\propto \int_{-\infty}^{+\infty} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} \\
&\propto \int_{-\infty}^{0} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} + \int_{0}^{+\infty} q\left(\delta_{ij} \mid U_{ij}\right) q\left(U_{ij}\right) dU_{ij} \\
&\propto \int_{-\infty}^{0} 1\left(\delta_{ij}=0\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \cdot q\left(U_{ij}\right) dU_{ij} \\
&\propto 1\left(\delta_{ij}=0\right) \cdot p\left(y_i^* \mid \alpha, \boldsymbol{\delta}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{X}_i, \boldsymbol{Z}_i, \boldsymbol{w}_i\right) \\
&\quad \cdot \int_{-\infty}^{0} \frac{B}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(U_{ij}-m_{U_{ij}}\right)^2\right\} dU_{ij} \\
&\propto \frac{B\cdot\Phi\left(-m_{U_{ij}}\right)}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \cdot (1-\delta_{ij}) \\
&\quad \cdot \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[y_i^* - \left(\alpha + \sum_{j'\neq j}^{m_i} \delta_{ij'}(1-w_{ij'})\boldsymbol{x}_{ij'}\boldsymbol{\beta} + \sum_{j'\neq j}^{m_i} \delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\right]^2\right\}.
\end{aligned}
$$

Therefore, we have

$$q\left(\delta_{ij}\right) \propto \begin{cases} \dfrac{A\left(1-\Phi\left(-m_{U_{ij}}\right)\right)}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)} \\ \cdot \dfrac{1}{\sqrt{2\pi}}\exp\left\{-\dfrac{1}{2}\left[y_i^* - \left(\alpha+(1-w_{ij})\boldsymbol{x}_{ij}\boldsymbol{\beta}+w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}+\sum\limits_{j'\neq j}^{m_i}\delta_{ij'}(1-w_{ij'})\boldsymbol{x}_{ij'}\boldsymbol{\beta}+\sum\limits_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\right]^2\right\}, & \text{if } \delta_{ij}=1; \\ \dfrac{B\cdot\Phi\left(-m_{U_{ij}}\right)}{A+(B-A)\cdot\Phi\left(-m_{U_{ij}}\right)}\cdot\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[y_i^*-\left(\alpha+\sum_{j'\neq j}^{m_i}\delta_{ij'}(1-w_{ij'})\boldsymbol{x}_{ij'}\boldsymbol{\beta}+\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\right]^2\right\}, & \text{if } \delta_{ij}=0. \end{cases}$$

- When $\delta_{ij}=1$,

$$\begin{aligned} \ln q^*\left(\delta_{ij}\right) = & \ln\frac{A^*\left(1-\Phi\left(-m_{U_{ij}}\right)\right)}{A^*+(B^*-A^*)\cdot\Phi\left(-m_{U_{ij}}\right)} - \ln\sqrt{2\pi} \\ & -\frac{1}{2}\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(y_i^*-\alpha-\sum_{j'\neq j}^{m_i}\delta_{ij'}\left(1-w_{ij'}\right)\boldsymbol{x}_{ij'}\boldsymbol{\beta}-\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)^2\right] \\ & +\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(y_i^*-\alpha-\sum_{j'\neq j}^{m_i}\delta_{ij'}\left(1-w_{ij'}\right)\boldsymbol{x}_{ij'}\boldsymbol{\beta}-\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\left(\left(1-w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta}+w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\right)\right] \\ & -\frac{1}{2}\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(1-w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^T+w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{z}_{ij}^T\right]. \end{aligned}$$

  Note that the A^* in $\ln\dfrac{A^*\left(1-\Phi\left(-m_{U_{ij}}\right)\right)}{A^*+(B^*-A^*)\cdot\Phi\left(-m_{U_{ij}}\right)}$ is the A computed in previous iteration.

- When $\delta_{ij}=0$,

$$\begin{aligned} \ln q^*\left(\delta_{ij}\right) = & \mathbb{E}_{q^*(-\delta_{ij})}\left[\ln\left(\frac{B^*\cdot\Phi\left(-m_{U_{ij}}\right)}{A^*+(B^*-A^*)\cdot\Phi\left(-m_{U_{ij}}\right)}\right)\right] \\ & +\mathbb{E}_{q^*(-\delta_{ij})}\left[\ln\left(\frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left[y_i^*-\left(\alpha+\sum_{j'\neq j}^{m_i}\delta_{ij'}\left(1-w_{ij'}\right)\boldsymbol{x}_{ij'}\boldsymbol{\beta}+\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\right]^2\right\}\right)\right] \\ = & \ln\frac{B^*\cdot\Phi\left(-m_{U_{ij}}\right)}{A^*+(B^*-A^*)\cdot\Phi\left(-m_{U_{ij}}\right)} \\ & -\ln\sqrt{2\pi}-\frac{1}{2}\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(y_i^*-\alpha-\sum_{j'\neq j}^{m_i}\delta_{ij'}\left(1-w_{ij'}\right)\boldsymbol{x}_{ij'}\boldsymbol{\beta}-\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)^2\right]. \end{aligned}$$

  Note that the B^* in $\ln\dfrac{B^*\cdot\Phi\left(-m_{U_{ij}}\right)}{A^*+(B^*-A^*)\cdot\Phi\left(-m_{U_{ij}}\right)}$ is the B computed in previous iteration.

Clearly $q^*(\delta_{ij})$ follows a Bernoulli distribution. By deleting terms appearing $\ln q^*(\delta_{ij})$ both when $\delta_{ij}=1$ and $\delta_{ij}=0$, we have

- When $\delta_{ij}=1$,

$$\begin{aligned}\ln q^*(\delta_{ij}) = \ln &\left(A^*\left(1-\Phi\left(-m_{U_{ij}}\right)\right)\right)\\ &+\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(y_i^*-\alpha-\sum_{j'\neq j}^{m_i}\delta_{ij'}\left(1-w_{ij'}\right)\boldsymbol{x}_{ij'}\boldsymbol{\beta}-\sum_{j'\neq j}^{m_i}\delta_{ij'}w_{ij'}\boldsymbol{z}_{ij'}\boldsymbol{\gamma}\right)\left(\left(1-w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta}+w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\right)\right]\\ &-\frac{1}{2}\mathbb{E}_{q^*(-\delta_{ij})}\left[\left(1-w_{ij}\right)\boldsymbol{x}_{ij}\boldsymbol{\beta}\boldsymbol{\beta}^T\boldsymbol{x}_{ij}^T+w_{ij}\boldsymbol{z}_{ij}\boldsymbol{\gamma}\boldsymbol{\gamma}^T\boldsymbol{z}_{ij}^T\right].\end{aligned}$$

- When $\delta_{ij}=0$,

$$\ln q^*(\delta_{ij}) = \ln\left(B^*\cdot\Phi\left(-m_{U_{ij}}\right)\right).$$

Therefore,

$$q^*(\delta_{ij}) \propto \begin{cases}\frac{A}{A+B}, & \text{if } \delta_{ij}=1;\\ \frac{B}{A+B}, & \text{if } \delta_{ij}=0.\end{cases}$$

where $A=\exp\left(\ln q^*(\delta_{ij})\right)$ when $\delta_{ij}=1$ and $B=\exp\left(\ln q^*(\delta_{ij})\right)$ when $\delta_{ij}=0$.

BIBLIOGRAPHY

[1] G. Wang, Y. Cheng, Y. Xia, Q. Ling and X. Wang, *A Bayesian semisupervised approach to keyword extraction with only positive and unlabeled data*, *INFORMS Journal on Computing* **35** (2023) 675–691. 1, 3, 12, 22, 23, 25, 60

[2] A. Hulth, *Improved automatic keyword extraction given more linguistic knowledge*, in *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pp. 216–223, 2003. 1, 15, 22

[3] C. Caragea, F. Bulgarov, A. Godea and S. D. Gollapalli, *Citation-enhanced keyphrase extraction from research papers: A supervised approach*, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1435–1446, 2014. 1

[4] M. Bordoloi, P. C. Chatterjee, S. K. Biswas and B. Purkayastha, *Keyword extraction using supervised cumulative textrank*, *Multimedia Tools and Applications* **79** (2020) 31467–31496. 1

[5] S. Beliga, A. Meštrović and S. Martinčić-Ipšić, *An overview of graph-based keyword extraction methods and approaches*, *Journal of Information and Organizational Sciences* **39** (2015) 1–20. 2

[6] R. Mihalcea and P. Tarau, *Textrank: Bringing order into text*, in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 404–411, 2004. 2

[7] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, *Computer Networks and ISDN Systems* **30** (1998) 107–117. 2

[8] A. Bougouin, F. Boudin and B. Daille, *Topicrank: Graph-based topic ranking for keyphrase extraction*, in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pp. 543–551, 2013. 2

[9] C. Florescu and C. Caragea, *A position-biased pagerank algorithm for keyphrase extraction*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017. 2

[10] F. Boudin, *Unsupervised keyphrase extraction with multipartite graphs*, in *16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pp. 667–672, 2018. 2

[11] K. Sparck Jones, *A statistical interpretation of term specificity and its application in retrieval*, *Journal of Documentation* **28** (1972) 11–21. 2

[12] S. R. El-Beltagy and A. Rafea, *KP-Miner: A keyphrase extraction system for English and Arabic documents*, *Information Systems* **34** (2009) 132–144. 2

[13] R. Campos, V. Mangaravite, A. Pasquali, A. M. Jorge, C. Nunes and A. Jatowt, *A text feature based automatic keyword extraction method for single documents*, in *European Conference on Information Retrieval*, pp. 684–691, Springer, 2018. 2

[14] D. Li, S. Li, W. Li, W. Wang and W. Qu, *A semi-supervised key phrase extraction approach: Learning from title phrases through a document semantic network*, in *Proceedings of the ACL 2010 Conference Short Papers*, pp. 296–300, 2010. 3, 6

[15] D. Zhou, O. Bousquet, T. Lal, J. Weston and B. Schölkopf, *Learning with local and global consistency*, *Advances in Neural Information Processing Systems* **16** (2003) . 3

[16] H. Ye and L. Wang, *Semi-supervised learning for neural keyphrase generation*, in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4142–4153, 2018. 3

[17] F. C. Jonathan and O. Karnalim, *Semi-supervised keyphrase extraction on scientific article using fact-based sentiment*, *TELKOMNIKA (Telecommunication Computing Electronics and Control)* **16** (2018) 1771–1778. 3

[18] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, New York, NY, 2006. 4, 6, 9, 33

[19] D. M. Blei, A. Kucukelbir and J. D. McAuliffe, *Variational inference: A review for statisticians*, *Journal of the American Statistical Association* **112** (2017) 859–877. 4, 33

[20] J. H. Albert and S. Chib, *Bayesian analysis of binary and polychotomous response data*, *Journal of the American Statistical Association* **88** (1993) 669–679. 4, 7, 32

[21] G. Parisi, *Statistical Field Theory*. Addison-Wesley, Boston, MA, USA, 1988. 7

[22] D. M. Blei and M. I. Jordan, *Variational inference for Dirichlet process mixtures*, *Bayesian Analysis* **1** (2006) 121–144. 7

[23] M. A. Newton, A. Noueiry, D. Sarkar and P. Ahlquist, *Detecting differential gene expression with a semiparametric hierarchical mixture method*, *Biostatistics* **5** (2004) 155–176. 10

[24] R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. 12

[25] F. Boudin, *PKE: an open source Python-based keyphrase extraction toolkit*, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pp. 69–73, 2016. 12

[26] A. R. Aronson, O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson et al., *The nlm indexing initiative.*, in *Proceedings of the AMIA Symposium*, p. 17, American Medical Informatics Association, 2000. 14, 15

[27] S. N. Kim, O. Medelyan, M.-Y. Kan and T. Baldwin, *SemEval-2010 Task 5 : Automatic keyphrase extraction from scientific articles*, in *Proceedings of the 5th International Workshop on Semantic Evaluation* (K. Erk and C. Strapparava, eds.), (Uppsala, Sweden), pp. 21–26, Association for Computational Linguistics, July, 2010. 15, 17

[28] N. G. Polson, J. G. Scott and J. Windle, *Bayesian inference for logistic models using Pólya–Gamma latent variables*, *Journal of the American Statistical Association* **108** (2013) 1339–1349. 25

[29] M. D. Hoffman, D. M. Blei, C. Wang and J. Paisley, *Stochastic variational inference*, *Journal of Machine Learning Research* **14** (2013) 1303–1347. 25

[30] J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith et al., *Non-centered parameterisations for hierarchical models and data augmentation*, in *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, vol. 307, Oxford University Press, USA, 2003. 25

[31] O. Papaspiliopoulos, G. O. Roberts and M. Sköld, *A general framework for the parametrization of hierarchical models*, *Statistical Science* (2007) 59–73. 25

[32] L. S. Tan, *Use of model reparametrization to improve variational Bayes*, *Journal of the Royal Statistical Society Series B: Statistical Methodology* **83** (2021) 30–57. 25

[33] L. S. Tan and D. J. Nott, *Variational inference for generalized linear mixed models using partially noncentered parametrizations*, *Statistical Science* **28** (2013) 168–188. 25

[34] S. M. Ali and S. D. Silvey, *A general class of coefficients of divergence of one distribution from another*, *Journal of the Royal Statistical Society: Series B (Methodological)* **28** (1966) 131–142. 26

[35] R. Bamler, C. Zhang, M. Opper and S. Mandt, *Perturbative black box variational inference*, *Advances in Neural Information Processing Systems* **30** (2017) . 26

[36] T. G. Dietterich, R. H. Lathrop and T. Lozano-Pérez, *Solving the multiple instance problem with axis-parallel rectangles*, *Artificial Intelligence* **89** (1997) 31–71. 27

[37] G. Quellec, G. Cazuguel, B. Cochener and M. Lamard, *Multiple-instance learning for medical image and video analysis*, *IEEE Reviews in Biomedical Engineering* **10** (2017) 213–234. 27

[38] J. Wu, Y. Yu, C. Huang and K. Yu, *Deep multiple instance learning for image classification and auto-annotation*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3460–3469, 2015. 27

[39] R. Bunescu and R. Mooney, *Learning to extract relations from the web using minimal supervision*, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 576–583, 2007. 27

[40] S. Angelidis and M. Lapata, *Multiple instance learning networks for fine-grained sentiment analysis*, *Transactions of the Association for Computational Linguistics* **6** (2018) 17–31. 27

[41] D. Pathak, E. Shelhamer, J. Long and T. Darrell, *Fully convolutional multi-class multiple instance learning*, *arXiv preprint arXiv:1412.7144* (2014) . 27

[42] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic and S. Vucetic, *Aerosol optical depth prediction from satellite observations by multiple instance regression*, in *Proceedings of the 2008 SIAM International Conference on Data Mining*, pp. 165–176, SIAM, 2008. 27

[43] S. Park, X. Wang, J. Lim, G. Xiao, T. Lu and T. Wang, *Bayesian multiple instance regression for modeling immunogenic neoantigens*, *Statistical Methods in Medical Research* **29** (2020) 3032–3047. 27

[44] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman and K. P. Bennett, *Fast bundle algorithm for multiple-instance learning*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **34** (2011) 1068–1079. 27

[45] D. Zhang, F. Wang, L. Si and T. Li, *Maximum margin multiple instance clustering with applications to image and text clustering*, *IEEE Transactions on Neural Networks* **22** (2011) 739–751. 27

[46] J. Amores, *Multiple instance classification: Review, taxonomy and comparative study*, *Artificial Intelligence* **201** (2013) 81–105. 27, 28, 38

[47] M.-A. Carbonneau, V. Cheplygina, E. Granger and G. Gagnon, *Multiple instance learning: A survey of problem characteristics and applications*, *Pattern Recognition* **77** (2018) 329–353. 28, 38

[48] D. Xiong, Z. Zhang, T. Wang and X. Wang, *A comparative study of multiple instance learning methods for cancer detection using t-cell receptor sequences*, *Computational and Structural Biotechnology Journal* **19** (2021) 3255–3268. 28

[49] P. Y. Chen, C. C. Chen, C. H. Yang, S. M. Chang and K. J. Lee, *Milr: Multiple-instance logistic regression with lasso penalty*, *R Journal* **9** (2017) 446–457. 28, 38

[50] D. Xiong, S. Park, J. Lim, T. Wang and X. Wang, *Bayesian multiple instance classification based on hierarchical probit regression*, *The Annals of Applied Statistics* **18** (2024) 80–99. 28, 32, 38, 45, 46

[51] S. Ray and D. Page, *Multiple instance regression*, in *Proceedings of the 18th International Conference on Machine Learning*, pp. 425–432, 2001. 28

[52] M. Dundar, B. Krishnapuram, R. Rao and G. Fung, *Multiple instance learning for computer aided diagnosis*, *Advances in Neural Information Processing Systems* **19** (2006) . 28

[53] Y. Ma, Z. Xiang, Q. Du and W. Fan, *Effects of user-provided photos on hotel review helpfulness: An analytical approach with deep leaning*, *International Journal of Hospitality Management* **71** (2018) 120–131. 29, 43, 46

[54] I. K. S. Al-Tameemi, M.-R. Feizi-Derakhshi, S. Pashazadeh and M. Asadpour, *Interpretable multimodal sentiment classification using deep multi-view attentive network of image and text data*, *IEEE Access* (2023) . 29, 46

[55] B. Xu, C. Lu, Y. Guo and J. Wang, *Discriminative multi-modality speech recognition*, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14433–14442, 2020. 29, 46

[56] X. Tang, J. Zhang, Y. He, X. Zhang, Z. Lin, S. Partarrieu et al., *Explainable multi-task learning for multi-modality biological data analysis*, *Nature Communications* **14** (2023) 2546. 29, 46

[57] L. Li, W. Ding, L. Huang, X. Zhuang and V. Grau, *Multi-modality cardiac image computing: A survey*, *Medical Image Analysis* (2023) 102869. 29

[58] M. Sahasrabudhe, P. Sujobert, E. I. Zacharaki, E. Maurin, B. Grange, L. Jallades et al., *Deep multi-instance learning using multi-modal data for diagnosis of lymphocytosis*, *IEEE Journal of Biomedical and Health Informatics* **25** (2020) 2125–2136. 29, 39, 46

[59] X. Li, Y. Zhou, J. Wang, H. Lin, J. Zhao, D. Ding et al., *Multi-modal multi-instance learning for retinal disease recognition*, in *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 2474–2482, 2021. 29, 39

[60] D. Durante and T. Rigon, *Conditionally conjugate mean-field variational bayes for logistic models*, *Statistical Science* **34** (2019) 472–485. 33

[61] Y. Zhang and Y. Yang, *Bayesian model selection via mean-field variational approximation*, *Journal of the Royal Statistical Society Series B: Statistical Methodology* (2024) qkad164. 33

[62] Q. Zhang and S. Goldman, *EM-DD: An improved multiple-instance learning technique*, *Advances in Neural Information Processing Systems* **14** (2001) . 38

[63] S. Andrews, I. Tsochantaridis and T. Hofmann, *Support vector machines for multiple-instance learning*, *Advances in Neural Information Processing Systems* **15** (2002) . 38

[64] S. Ray and M. Craven, *Supervised versus multiple instance learning: An empirical comparison*, in *Proceedings of the 22nd International Conference on Machine Learning*, pp. 697–704, 2005. 38

[65] B. Babenko, P. Dollár, Z. Tu and S. Belongie, *Simultaneous learning and alignment: Multi-instance and multi-pose learning*, in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008. 38

[66] J. Wang and J.-D. Zucker, *Solving the multiple-instance problem: A lazy learning approach*, in *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1119–1126, 2000. 38

[67] T. Gärtner, P. A. Flach, A. Kowalczyk and A. J. Smola, *Multi-instance kernels*, in *Proceedings of the 19th International Conference on Machine Learning*, pp. 179–186, 2002. 38

[68] J. Zhang, M. Marszałek, S. Lazebnik and C. Schmid, *Local features and kernels for classification of texture and object categories: A comprehensive study*, *International Journal of Computer Vision* **73** (2007) 213–238. 38

[69] Z.-H. Zhou, Y.-Y. Sun and Y.-F. Li, *Multi-instance learning by treating instances as non-iid samples*, in *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1249–1256, 2009. 38

[70] V. Cheplygina, D. M. Tax and M. Loog, *Multiple instance learning with bag dissimilarities*, *Pattern Recognition* **48** (2015) 264–275. 38

[71] Y. Chen, J. Bi and J. Z. Wang, *MILES: Multiple-instance learning via embedded instance selection*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **28** (2006) 1931–1947. 38

[72] Z.-H. Zhou and M.-L. Zhang, *Solving multi-instance problems with classifier ensemble based on constructive clustering*, *Knowledge and Information Systems* **11** (2007) 155–170. 38

[73] E.-J. Lee and S. Y. Shin, *When do consumers buy online product reviews? effects of review quality, product type, and reviewer's photo*, *Computers in Human Behavior* **31** (2014) 356–366. 41

[74] Y. Li and Y. Xie, *Is a picture worth a thousand words? an empirical study of image content and social media engagement*, *Journal of Marketing Research* **57** (2020) 1–19. 43

[75] Q.-T. Truong and H. W. Lauw, *Vistanet: Visual aspect attention network for multimodal sentiment analysis*, in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 305–312, 2019. 43

[76] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal et al., *Learning transferable visual models from natural language supervision*, in *International Conference on Machine Learning*, pp. 8748–8763, PMLR, 2021. 43

[77] D. P. Kingma and M. Welling, *Auto-encoding variational Bayes*, *arXiv preprint arXiv:1312.6114* (2013) . 43