Supplementary Material

1. Summary of Notation

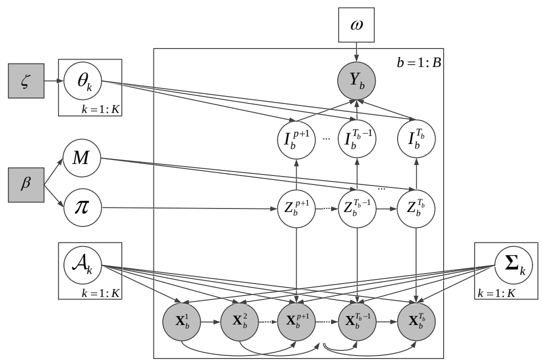

Figure 1. The graphical model representation of the ARHMM-MIL model

Symbol	Size	Description
Variables		
Y_b	Scalar	The label of bag b
I_b^t	Scalar	tth instance label in bag b
$\mathbf{I}_b$	$1 \times (T_b - p)$	All the instance labels in bag b
$\mathbf{X}_b^t$	$d \times 1$	The tth observation in bag b
$\mathbf{X}_b$	$d \times T_b$	All the observations in bag b
$\mathbf{X}_b^{(t-p):(t-1)}$	$d \times p$	p observations prior to the tth observation in bag b
Z_b^t	Scalar	Cluster membership of instance t in bag b
Model Parameters		
ω	Scalar	Parameter of the softmax function
$\boldsymbol{\beta}$	$K \times 1$	Dirichlet prior parameter for the cluster membership
$\boldsymbol{\pi}$	$K \times 1$	Probability for initial cluster assignment
$\mathbf{M}$	$K \times K$	Transition matrix, $M_{ij} = P(Z_b^t = j \mid Z_b^{t-1} = i, \mathbf{M})$
$\mathbf{A}_{k0}$	$d \times 1$	Intercept term of the kth AR cluster
$\mathbf{A}_{kj}$	$d \times d$	Coefficient of the jth order of the kth AR cluster
$\boldsymbol{\Sigma}_k$	$d \times d$	Covariance matrix of the kth AR cluster
$\boldsymbol{\zeta}$	2×1	Beta prior for each cluster
θ_k	Scalar	The Bernoulli parameter for the kth cluster
Dimensions		
B		Number of bags
T_b		Length of time series in bag b
K		Number of mixture components
p		Order of each auto-regressive process
d		Time series dimension

Table 1. A summary of the notation used.

2. E-step Derivation

The complete-data log-likelihood is as follows:

$$
\begin{aligned}
\ell(\boldsymbol{\Theta}) &= \sum_{b=1}^{B} \log P(\mathbf{Z}_b, \mathbf{I}_b, \mathbf{X}_b, Y_b|\boldsymbol{\Theta}) \\
&= \sum_{b=1}^{B} \Big[\log P(Y_b|\mathbf{I}_b, \omega) + \log P(\mathbf{I}_b|\mathbf{Z}_b, \boldsymbol{\theta}) + \log P(\mathbf{Z}_b|\mathbf{M}, \boldsymbol{\pi}) + \log P(\mathbf{X}_b|\mathbf{Z}_b, \boldsymbol{\mathcal{A}}, \boldsymbol{\Sigma}) \Big] \\
&= \sum_{b=1}^{B} \Big[\log P(Y_b|\mathbf{I}_b, \omega) + \sum_{t=p+1}^{T_b} \log P(I_b^t|Z_b^t, \boldsymbol{\theta}) + \log P(Z_b^{p+1}|\boldsymbol{\pi}) + \sum_{t=p+2}^{T_b} \log P(Z_b^t|Z_b^{t-1}, \mathbf{M}) \\
&\quad + \sum_{t=p+1}^{T_b} \log P(\mathbf{X}_b^t|\mathbf{X}_b^{(t-p):(t-1)}, Z_b^t, \boldsymbol{\mathcal{A}}, \boldsymbol{\Sigma}) \Big]
\end{aligned} \tag{1}
$$

In order to form the auxiliary function $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}')$, we take the expected value of the complete data log-likelihood under the distribution $P(\mathbf{Z}, \mathbf{I}|\mathbf{X}, \mathbf{Y}, \boldsymbol{\Theta}')$.

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}') &= E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\sum_{b=1}^{B} \sum_{\mathbf{I}} \mathbb{I}(\mathbf{I}_b = \mathbf{I}) \Big(Y_b \log P(Y_b = 1|\mathbf{I}_b, \omega) + (1 - Y_b) \log P(Y_b = 0|\mathbf{I}_b, \omega) \Big) \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{l=0}^{1} \sum_{j=1}^{K} \mathbb{I}(Z_b^t = j, I_b^t = l) \log P(\mathbf{I}_b^t = l|Z_b^t = j, \boldsymbol{\theta}) \\
&\quad + \sum_{b=1}^{B} \sum_{j=1}^{K} \mathbb{I}(Z_b^{p+1} = j) \log P(Z_b^{p+1} = j|\boldsymbol{\pi}) \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+2}^{T_b} \sum_{j=1}^{K} \sum_{i=1}^{K} \mathbb{I}(Z_b^t = j, Z_b^{t-1} = i) \log P(Z_b^t = j|Z_b^{t-1} = i, \mathbf{M}) \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{j=1}^{K} \mathbb{I}(Z_b^t = j) \log P(\mathbf{X}_b^t|\mathbf{X}_b^{(t-p):(t-1)}, Z_b^t = j, \boldsymbol{\mathcal{A}}, \boldsymbol{\Sigma}) \Big] \\
&= \sum_{b=1}^{B} \sum_{\mathbf{I}} E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\mathbb{I}(\mathbf{I}_b = \mathbf{I}) \Big(Y_b \log P(Y_b = 1|\mathbf{I}_b, \omega) + (1 - Y_b) \log P(Y_b = 0|\mathbf{I}_b, \omega) \Big) \Big] \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{l=0}^{1} \sum_{j=1}^{K} E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\mathbb{I}(Z_b^t = j, I_b^t = l) \log P(\mathbf{I}_b^t = l|Z_b^t = j, \boldsymbol{\theta}) \Big] \\
&\quad + \sum_{b=1}^{B} \sum_{j=1}^{K} E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\mathbb{I}(Z_b^{p+1} = j) \log P(Z_b^{p+1} = j|\boldsymbol{\pi}) \Big] \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+2}^{T_b} \sum_{j=1}^{K} \sum_{i=1}^{K} E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\mathbb{I}(Z_b^t = j, Z_b^{t-1} = i) \log P(Z_b^t = j|Z_b^{t-1} = i, \mathbf{M}) \Big] \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{j=1}^{K} E_{P(\mathbf{Z},\mathbf{I}|\mathbf{X},\mathbf{Y},\boldsymbol{\Theta}')} \Big[\mathbb{I}(Z_b^t = j) \log P(\mathbf{X}_b^t|\mathbf{X}_b^{(t-p):(t-1)}, Z_b^t = j, \boldsymbol{\mathcal{A}}, \boldsymbol{\Sigma}) \Big] \\
&= \sum_{b=1}^{B} \sum_{\mathbf{I}} P(\mathbf{I}_b = \mathbf{I}|\mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \Big(Y_b \log P(Y_b = 1|\mathbf{I}_b, \omega) + (1 - Y_b) \log P(Y_b = 0|\mathbf{I}_b, \omega) \Big) \\
&\quad + \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{l=0}^{1} \sum_{j=1}^{K} P(I_b^t = l, Z_b^t = j|\mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log P(\mathbf{I}_b^t = l|Z_b^t = j, \boldsymbol{\theta})
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{b=1}^{B} \sum_{j=1}^{K} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \mathbf{\Theta}') \log P(Z_b^{p+1} = j | \boldsymbol{\pi}) \\
&+ \sum_{b=1}^{B} \sum_{t=p+2}^{T_b} \sum_{j=1}^{K} \sum_{i=1}^{K} P(Z_b^t = j, Z_b^{t-1} = i | \mathbf{X}_b, Y_b, \mathbf{\Theta}') \log P(Z_b^t = j | Z_b^{t-1} = i, \mathbf{M}) \\
&+ \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{j=1}^{K} P(Z_b^t = j | \mathbf{X}_b, Y_b, \mathbf{\Theta}') \log P(\mathbf{X}_b^t | \mathbf{X}_b^{(t-p):(t-1)}, Z_b^t = j, \boldsymbol{\mathcal{A}}, \mathbf{\Sigma})
\end{aligned} \tag{2}
$$

2.1. Message passing for a generalized version of a chain model

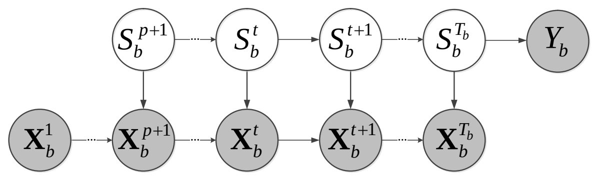

Figure 2. A simplified graphical model by representing the (N, Z) variables pair as a supernode S.

Before we apply the message passing to the proposed model, we convert the (N, Z) variables pair to a supernode S to introduce the message-passing algorithm on the simplified graphical model in Fig. 2. To simplify further derivations, we omit the conditioning on the parameter set $\mathbf{\Theta}'$.

In this approach we consider the following steps:

- A first pass in which a **forward message** is computed while traversing the graphical model from left to right; specifically, the forward message is initialized at $t = p+1$ by computing $\alpha_b^q(p+1) = P(\mathbf{X}_b^{1:p+1}, S_b^t = q)$ and for $t = p+2, \ldots, T_b$ is computed recursively using

$$
\begin{aligned}
\alpha_b^q(t) &= P(\mathbf{X}_b^{1:t}, S_b^t = q) \\
&= \sum_p P(\mathbf{X}_b^{1:t-1}, S_b^{t-1} = p) P(S_b^t = q | S_b^{t-1} = p) P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, S_b^t = q) \\
&= \sum_p \alpha_b^p(t-1) P(S_b^t = q | S_b^{t-1} = p) P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, S_b^t = q).
\end{aligned}
$$

- A second pass in which a **backward message** is computed while traversing the graphical model from right to left; the backward message is initialized at $t = T_b$ by setting $\beta_b^q(T_b) = P(Y_b | \mathbf{X}_b^{1:T_b}, S_b^{T_b} = q)$ and for $t = T_b - 1, \ldots, p+1$ is computed recursively by

$$
\begin{aligned}
\beta_b^q(t) &= P(Y_b, \mathbf{X}_b^{t+1:T_b} | \mathbf{X}_b^{1:t}, S_b^t = q) \\
&= \sum_r P(S_b^{t+1} = r | S_b^t = q) P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, S_b^{t+1} = r) P(Y_b, \mathbf{X}_b^{t+2:T_b} | \mathbf{X}_b^{1:t+1}, S_b^{t+1} = r) \\
&= \sum_r P(S_b^{t+1} = r | S_b^t = q) P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, S_b^{t+1} = r) \beta_b^r(t+1)
\end{aligned}
$$

- Finally the messages are used to form the pairwise probability for (S_b^t, S_b^{t-1}) conditioned on the observed nodes $\mathbf{X}_b^1, \ldots, \mathbf{X}_b^T$ and Y_b. The E-step calculation of the proposed model necessitates the probability of the form $P(S_b^t = q, S_b^{t-1} = r | \mathbf{X}_b, Y_b)$ given by

$$
P(S_b^t = q, S_b^{t-1} = r | \mathbf{X}_b, Y_b) = \frac{P(S_b^t = q, S_b^{t-1} = r, \mathbf{X}_b, Y_b)}{\sum_q \sum_r P(S_b^t = q, S_b^{t-1} = r, \mathbf{X}_b, Y_b)}. \tag{3}
$$

We focus on the joint distribution in the numerator of (3) since the denominator can be computed by marginalizing out S_b^t and S_b^{t-1} in the joint distribution. The numerator of (3) can be written in term of the forward message, the backward message, the state transition probability, and the observation model probability as

$$\begin{aligned}
&P(S_b^t = q, S_b^{t-1} = r, \mathbf{X}_b, Y_b) \\
&= P(Y_b, \mathbf{X}_b^{t+1:T_b} | \mathbf{X}_b^{1:t}, S_b^t = q) P(S_b^t = q | S_b^{t-1} = r) P(\mathbf{X}_b^t | \mathbf{X}_b^{t-p:t-1}, S_b^t = q) P(\mathbf{X}_b^{1:t-1}, S_b^{t-1} = r) \\
&= \beta_b^q(t) P(S_b^t = q | S_b^{t-1} = r) P(\mathbf{X}_b^t | \mathbf{X}_b^{t-p:t-1}, S_b^t = q) \alpha_b^r(t-1)
\end{aligned} \tag{4}$$

This approach provides the framework for computing the E-step probability terms after expanding the node S to (N_b^t, Z_b^t) as follows.

2.2. Message passing for the chain based on (N_b^t, Z_b^t)

In this section, we provide more detailed intermediate steps for the forward/backward message passing based on (N_b^t, Z_b^t). Recall that we use a simplified graphical model to represent the structure with in a single bag as in Fig. 3. We use the q and r to compactly denote $q = (q_N, q_Z)$ and $r = (r_N, r_Z)$ for $t = p+1, \cdots, T_b$.

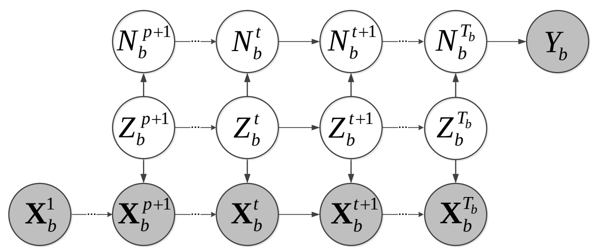

Figure 3. A graphical model representing a bag in Fig. 1 with the instance label I replaced by a counting variable N.

2.2.1. Forward Message

We assume the first p observations $\mathbf{X}_1, \ldots, \mathbf{X}_p$ follow a joint distribution $P(\mathbf{X}_b^{1:p})$ that is independent of any proposed model parameters. The forward message starts with $t = p+1$, so it is clear that $N_b^{p+1} = I_b^{p+1}$. Hence, the forward message at $t = p+1$ is initialized by

$$\begin{aligned}
\alpha_b^{q_N, q_Z}(p+1) &= P(\mathbf{X}_b^{1:p+1}, N_b^{p+1} = q_N, Z_b^{p+1} = q_Z) \\
&= P(\mathbf{X}_b^{p+1}, N_b^{p+1} = q_N, Z_b^{p+1} = q_Z | \mathbf{X}_b^{1:p}) P(\mathbf{X}_b^{1:p}) \\
&= P(\mathbf{X}_b^{p+1} | \mathbf{X}_b^{1:p}, Z_b^{p+1} = q_Z) P(N_b^{p+1} = q_N | Z_b^{p+1} = q_Z) P(Z_b^{p+1} = q_Z) P(\mathbf{X}_b^{1:p}) \\
&= P(\mathbf{X}_b^{p+1} | \mathbf{X}_b^{1:p}, Z_b^{p+1} = q_Z) P(I_b^{p+1} = q_N | Z_b^{p+1} = q_Z) P(Z_b^{p+1} = q_Z) P(\mathbf{X}_b^{1:p}) \\
&= A_b^{q_Z}(p+1) (\theta_{q_Z})^{q_N} (1 - \theta_{q_Z})^{(1-q_N)} \pi_{q_Z} P(\mathbf{X}_b^{1:p})
\end{aligned} \tag{5}$$

for $q_N \in \{0, 1\}$ and $q_Z \in \{1, \ldots, K\}$, where $A_b^{q_Z}(p+1) = P(\mathbf{X}_b^{p+1} | \mathbf{X}_b^{1:p}, Z_b^{p+1} = q_Z)$, $\pi_{q_Z} = P(Z_b^{p+1} = q_Z)$ and $\theta_{q_Z} = P(I_b^{p+1} = 1 | Z_b^{p+1} = q_Z)$. Then, the forward message is recursively computed for $t = p+2, \ldots, T_b$ using

$$\begin{aligned}
\alpha_b^{q_N, q_Z}(t) &= P(\mathbf{X}_b^{1:t}, N_b^t = q_N, Z_b^t = q_Z) \\
&= P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z) P(N_b^t = q_N, Z_b^t = q_Z, \mathbf{X}_b^{1:t-1}) \\
&= P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z) \sum_{r_N} \sum_{r_Z} P(N_b^t = q_N, Z_b^t = q_Z, N_b^{t-1} = r_N, Z_b^{t-1} = r_Z, \mathbf{X}_b^{1:t-1}) \\
&= P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z) \sum_{r_N} \sum_{r_Z} P(N_b^t = q_N | Z_b^t = q_Z, N_b^{t-1} = r_N) \\
&\quad \cdot P(Z_b^t = q_z | Z_b^{t-1} = r_Z) P(\mathbf{X}_b^{1:t-1}, N_b^{t-1} = r_N, Z_b^{t-1} = r_Z) \\
&= P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z) \sum_{r_N} \sum_{r_Z} P(N_b^t = q_N | Z_b^t = q_I, N_b^{t-1} = r_N)
\end{aligned}$$

$$\cdot P(Z_b^t = q_z | Z_b^{t-1} = r_Z) \alpha_b^{r_N, r_Z}(t-1)$$

$$= A_b^{q_Z}(t) \sum_{r_Z} \sum_{r_N} \Big(\mathbb{I}(r_N = q_N)(1 - \theta_{q_Z}) + \mathbb{I}(r_N = q_N + 1)\theta_{q_Z} \Big) M_{r_Z, q_Z} \alpha_b^{r_N, r_Z}(t-1) \tag{6}$$

where $M_{r_Z,q_Z} = P(Z_b^t = q_Z | Z_b^{t-1} = r_Z)$, $A_b^{q_Z}(t) = P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z)$ and $\theta_{q_Z} = P(I_b^t = 1 | Z_b^t = q_Z)$.

2.2.2. Backward Message

The backward message is initialized at $t = T_b$ with

$$\beta_b^{q_N, q_Z}(T_b) = P(Y_b | \mathbf{X}_b^{1:T_b}, Z_b^{T_b} = q_Z, N_b^{T_b} = q_N) = P(Y_b | N_b^{T_b} = q_N) \tag{7}$$

for all $q_Z \in \{1, \dots, K\}$, where $P(Y_b | N_b^{T_b} = q_N)$ is computed using the positive bag probability (29). Then, the backward message is recursively computed for $t = T_b - 1, \dots, p+1$ using

$$\begin{aligned}
\beta_b^{q_N, q_Z}(t) &= P(Y_b, \mathbf{X}_b^{t+1:T_b} | \mathbf{X}_b^{1:t}, N_b^t = q_N, Z_b^t = q_Z) \\
&= \sum_{r_N} \sum_{r_Z} P(Y_b, \mathbf{X}_b^{t+1:T_b}, N_b^{t+1} = r_N, Z_b^{t+1} = r_Z | \mathbf{X}_b^{1:t}, N_b^t = q_N, Z_b^t = q_Z) \\
&= \sum_{r_N} \sum_{r_Z} P(N_b^{t+1} = r_N | Z_b^{t+1} = r_Z, N_b^t = q_N) P(Z_b^{t+1} = r_Z | Z_b^t = q_Z) \\
&\quad \cdot P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, Z_b^{t+1} = r_Z) P(Y_b, \mathbf{X}_b^{t+2:T_b} | \mathbf{X}_b^{1:t+1}, N_b^{t+1} = r_N, Z_b^{t+1} = r_Z) \\
&= \sum_{r_N} \sum_{r_Z} P(N_b^{t+1} = r_N | Z_b^{t+1} = r_Z, N_b^t = q_N) P(Z_b^{t+1} = r_Z | Z_b^t = q_Z) \\
&\quad \cdot P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, Z_b^{t+1} = r_Z) \beta_b^{r_N, r_Z}(t+1) \\
&= \sum_{r_Z} P(Z_b^{t+1} = r_Z | Z_b^t = q_Z) P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, Z_b^{t+1} = r_Z) \\
&\quad \cdot \sum_{r_N} P(N_b^{t+1} = r_N | Z_b^{t+1} = r_Z, N_b^t = q_N) \beta_b^{r_N, r_Z}(t+1) \\
&= \sum_{r_Z} M_{q_Z, r_Z} A_b^{r_Z}(t+1) \sum_{r_N} \Big(\mathbb{I}(r_N = q_N)(1 - \theta_{r_Z}) + \mathbb{I}(r_N = 1 + q_N)\theta_{r_Z} \Big) \beta_b^{r_N, r_Z}(t+1)
\end{aligned}$$

where $M_{q_Z,r_Z} = P(Z_b^{t+1} = r_Z | Z_b^t = q_Z)$, $A_b^{r_Z}(t+1) = P(\mathbf{X}_b^{t+1} | \mathbf{X}_b^{1:t}, Z_b^{t+1} = r_Z)$ and $\theta_{r_Z} = P(I_b^{t+1} = 1 | Z_b^{t+1} = r_Z)$ (as previously defined).

2.2.3. Expanding the pairwise state probability

We expand (4) by changing S_b^t to the pair (N_b^t, Z_b^t) where we can compute the pairwise state probability $P(N_b^t = q_N, Z_b^t = q_Z, N_b^{t-1} = r_N, Z_b^{t-1} = r_Z, \mathbf{X}_b, Y_b)$ using

$$\begin{aligned}
&P(N_b^t = q_N, Z_b^t = q_Z, N_b^{t-1} = r_N, Z_b^{t-1} = r_Z, \mathbf{X}_b, Y_b) \\
&= P(Y_b, \mathbf{X}_b^{t+1:T_b} | \mathbf{X}_b^{1:t}, N_b^t = q_N, Z_b^t = q_Z) P(N_b^t = q_N, Z_b^t = q_Z | N_b^{t-1} = r_N, Z_b^{t-1} = r_Z) \\
&\quad \cdot P(\mathbf{X}_b^t | N_b^t = q_N, Z_b^t = q_Z, \mathbf{X}_b^{1:t-1}) P(\mathbf{X}_b^{1:t-1}, N_b^{t-1} = r_N, Z_b^{t-1} = r_Z) \\
&= \beta_b^{q_N, q_Z}(t) P(N_b^t = q_N, Z_b^t = q_Z | N_b^{t-1} = r_N, Z_b^{t-1} = r_Z) P(\mathbf{X}_b^t | N_b^t = q_N, Z_b^t = q_Z, \mathbf{X}_b^{1:t-1}) \alpha_b^{r_N, r_Z}(t-1) \\
&= \beta_b^{q_N, q_Z}(t) P(N_b^t = q_N | Z_b^t = q_Z, N_b^{t-1} = r_N) P(Z_b^t = q_Z | Z_b^{t-1} = r_Z) P(\mathbf{X}_b^t | Z_b^t = q_Z, \mathbf{X}_b^{1:t-1}) \alpha_b^{r_N, r_Z}(t-1) \\
&= \beta_b^{q_N, q_Z}(t) P(I_b^t = q_N - r_N | Z_b^t = q_Z) P(Z_b^t = q_Z | Z_b^{t-1} = r_Z) P(\mathbf{X}_b^t | Z_b^t = q_Z, \mathbf{X}_b^{1:t-1}) \alpha_b^{r_N, r_Z}(t-1) \\
&= \beta_b^{q_N, q_Z}(t) \theta_{q_Z}^{\mathbb{I}(q_N = r_N)} (1 - \theta_{q_Z})^{\mathbb{I}(q_N = r_N - 1)} M_{r_Z, q_Z} A_b^{q_Z}(t) \alpha_b^{r_N, r_Z}(t-1).
\end{aligned} \tag{8}$$

where $M_{r_Z,q_Z} = P(Z_b^t = q_Z | Z_b^{t-1} = r_Z)$, $A_b^{q_Z}(t) = P(\mathbf{X}_b^t | \mathbf{X}_b^{1:t-1}, Z_b^t = q_Z)$ and $\theta_{q_Z} = P(I_b^t = 1 | Z_b^t = q_Z)$.

3. M-Step Derivation

In this section, we derive the equations for the M-step for all the parameters of the ARHMM-MIL model. Under the maximum-a-posterior(MAP) framework, the objective function for the M-step is

$$
\begin{aligned}
&\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}') + \log P(\boldsymbol{\Theta}) \\
&= \sum_{b=1}^{B} \sum_{\mathbf{I}} P(\mathbf{I}_b = \mathbf{I} | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \Big(Y_b \log P(Y_b = 1 | \mathbf{I}_b, \omega) + (1 - Y_b) \log P(Y_b = 0 | \mathbf{I}_b, \omega) \Big) \\
&\quad + \sum_{j=1}^{K} \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} \sum_{l=0}^{1} P(I_b^t = l, Z_b^t = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log P(\mathbf{I}_b^t = l | Z_b^t = j, \boldsymbol{\theta}) + \log P(\theta_j | \boldsymbol{\zeta}) \\
&\quad + \sum_{j=1}^{K} \sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log P(Z_b^{p+1} = j | \boldsymbol{\pi}) + \log P(\pi_j | \boldsymbol{\beta}) \\
&\quad + \sum_{j=1}^{K} \sum_{i=1}^{K} \sum_{b=1}^{B} \sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log P(Z_b^t = j | Z_b^{t-1} = i, \mathbf{M}) + \log P(M_{ij} | \boldsymbol{\beta}) \\
&\quad + \sum_{j=1}^{K} \sum_{b=1}^{B} \sum_{t=p+1}^{T_b} P(Z_b^t = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log P(\mathbf{X}_b^t | \mathbf{X}_b^{(t-p):(t-1)}, Z_b^t = j, \boldsymbol{\mathcal{A}}, \boldsymbol{\Sigma}) \qquad (9)
\end{aligned}
$$

In general, each equation in the M-step involves a maximum likelihood estimation problem. We simply take the derivative of the objective function (9) with respect to each parameter, set the derivative to zero and solve. We can efficiently estimate the parameter if its closed-form solution exists; otherwise, we apply gradient ascent.

3.1. Update the initial prior π of the hidden states Z

In order to update π_j, we collect the terms in (9) that involve π_j and also add a Lagrangian multiplier η for the constraint $\sum_{j=1}^{K} \pi_j = 1$. The resulting function is:

$$
L(\pi_j) = \sum_{j=1}^{K} \left(\sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') \log \pi_j + (\beta_j - 1) \log \pi_j \right) + \eta (\sum_{j=1}^{K} \pi_j - 1) \qquad (10)
$$

Taking derivative of $L(\pi_j)$ with respect to π_j and setting to 0 we get

$$
\frac{\sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') + \beta_j - 1}{\pi_j} + \eta = 0. \qquad (11)
$$

Setting the derivative to 0 and solving for π_j results in:

$$
\pi_j = \frac{- \sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') + \beta_j - 1}{\eta} \qquad (12)
$$

Set $\frac{\partial L(\pi_j)}{\partial \eta} = \sum_{j=1}^{K} \pi_j - 1 = 0$ we get $\sum_{j=1}^{K} \pi_j = 1$. If we sum up (12) over $j = 1, \ldots, K$, we get:

$$
\begin{aligned}
\sum_{j=1}^{K} \pi_j &= \sum_{j=1}^{K} \frac{- \sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') + \beta_j - 1}{\eta} \\
1 &= \sum_{j=1}^{K} \frac{- \sum_{b=1}^{B} P(Z_b^{p+1} = j | \mathbf{X}_b, Y_b, \boldsymbol{\Theta}') + \beta_j - 1}{\eta}
\end{aligned}
$$

$$\eta = -\sum_{j=1}^{K}\sum_{b=1}^{B} P(Z_b^{p+1} = j|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1 \tag{13}$$

Substitute (13) into (12), we get:

$$\pi_j = \frac{\sum_{b=1}^{B} P(Z_b^{p+1} = j|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{\sum_{j'=1}^{K}\sum_{b=1}^{B} P(Z_b^{p+1} = j'|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_{j'} - 1} \tag{14}$$

3.2. Update the transition matrix M of the hidden states

Since the rows in the hidden state transition matrix sum to 1, we need to include the constraints that $\sum_{j=1}^{K} M_{ij} = 1$ for $i = 1, \dots K$. Similar to (10), the Lagrangian of M_{ij} can be formulated by collecting all terms involving M_{ij} in (9) and introducing Lagrangian multipliers η_i for $i = 1, \dots, K$:

$$L(M_{ij}) = \sum_{j=1}^{K}\sum_{i=1}^{K}\left(\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') \log M_{ij} + (\beta_j - 1)\log M_{ij}\right) + \sum_{i=1}^{K}\eta_i(\sum_{j=1}^{K} M_{ij} - 1) \tag{15}$$

Taking the derivative of $L(M_{ij})$ with respect to M_{ij} and setting it equal to 0 we obtain:

$$\frac{\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{M_{ij}} + \eta_i = 0$$

$$\implies M_{ij} = \frac{-\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{\eta_i} \tag{16}$$

Set $\frac{\partial L(M_{ij})}{\partial \eta_i} = \sum_{j=1}^{K} M_{ij} = 0$ we get $\sum_{j=1}^{K} M_{ij} = 1$. If we sum up (16) over $j = 1, \dots, K$, we get:

$$\sum_{j=1}^{K} M_{ij} = \sum_{j=1}^{K}\frac{-\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{\eta_i}$$

$$\implies 1 = \frac{-\sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{\eta_i}$$

$$\implies \eta_i = -\sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1 \tag{17}$$

Substitute (17) into (16), we get the update equation for M_{ij}:

$$M_{ij} = \frac{\sum_{b=1}^{B}\sum_{t=p+2}^{T_b} P(Z_b^t = j, Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_j - 1}{\sum_{j'=1}^{K}\sum_{b=1}^{B}\sum_{t=p+2}^{T_b+} P(Z_b^t = j', Z_b^{t-1} = i|\mathbf{X}_b, Y_b, \mathbf{\Theta}') + \beta_{j'} - 1} \tag{18}$$

3.3. Update the covariance matrix $\mathbf{\Sigma}_j$ for each AR process

In order to simplify the notation, we denote that $\xi_{btj} = P(Z_b^t = j|\mathbf{X}_b, Y_b, \mathbf{\Theta}')$ is a $K \times K$ matrix for $t = p+1, \ldots, T_b$. The objective for estimating $\mathbf{\Sigma}_j$ can be written as:

$$L(\mathbf{\Sigma}_j) = \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}(\mathbf{X}_b^t - \mathbf{A}_{j0} - \sum_{k=1}^{p}\mathbf{A}_{jk}\mathbf{X}_b^{t-k})'\mathbf{\Sigma}_j^{-1}(\mathbf{X}_b^t - \mathbf{A}_{j0} - \sum_{k=1}^{p}\mathbf{A}_{jk}\mathbf{X}_b^{t-k}) - \frac{1}{2}\log|\mathbf{\Sigma}_j|\Bigg] \quad (19)$$

For convenience, denote $\boldsymbol{\mu}_{btj} = \mathbf{X}_b^t - (\mathbf{A}_{j0} + \sum_{k=1}^{p}\mathbf{A}_{jk}\mathbf{X}_b^{t-k})$.

$$\begin{aligned}
L(\mathbf{\Sigma}_j) &= \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}\boldsymbol{\mu}_{btj}{}'\mathbf{\Sigma}_j^{-1}\boldsymbol{\mu}_{btj} - \frac{1}{2}\log|\mathbf{\Sigma}_j|\Bigg] \\
&= \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}Tr(\boldsymbol{\mu}_{btj}{}'\mathbf{\Sigma}_j^{-1}\boldsymbol{\mu}_{btj}) - \frac{1}{2}\log|\mathbf{\Sigma}_j|\Bigg] \\
&= \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}Tr(\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}'\mathbf{\Sigma}_j^{-1}) - \frac{1}{2}\log|\mathbf{\Sigma}_j|\Bigg]
\end{aligned}$$

Using the fact that $|\mathbf{\Sigma}_j^{-1}|^{-1} = |\mathbf{\Sigma}_j|$,

$$L(\mathbf{\Sigma}_j) = \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}Tr(\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}'\mathbf{\Sigma}_j^{-1}) + \frac{1}{2}\log|\mathbf{\Sigma}_j^{-1}|\Bigg]$$

Now, instead of differentiating $L(\mathbf{\Sigma}_j)$ with respect to $\mathbf{\Sigma}_j$, we differentiate with respect to $\mathbf{\Sigma}_j^{-1}$. Setting the derivative to 0 and solving, we get:

$$\begin{aligned}
&\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}' + \frac{1}{2|\mathbf{\Sigma}_j^{-1}|}|\mathbf{\Sigma}_j^{-1}|\mathbf{\Sigma}_j\Bigg] = 0 \\
&\implies \sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[-\frac{1}{2}\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}' + \frac{1}{2}\mathbf{\Sigma}_j\Bigg] = 0 \\
&\implies \sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\Bigg[\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}' - \mathbf{\Sigma}_j\Bigg] = 0 \\
&\implies \sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}' = \sum_{i=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\sum_{l=0}^{1}\xi_{b,t,l}^{i,j}\mathbf{\Sigma}_j \\
&\implies \mathbf{\Sigma}_j = \frac{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}\boldsymbol{\mu}_{btj}\boldsymbol{\mu}_{btj}{}'}{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \xi_{btj}} \qquad (20)
\end{aligned}$$

3.4. Update the AR coefficients $\mathbf{A}_{j0}$ and $\mathbf{A}_{jk}$

In (21) below, we collect the terms in (9) that are related to the AR coefficients. Recall that $\mathbf{A}_{j0}$ is a $d \times 1$ vector and $\mathbf{A}_{jk}$ are matrices of size $d \times d$.

$$L(\mathbf{A}_{j0}, \mathbf{A}_{j1}, \ldots, \mathbf{A}_{jp}) =$$

$$\sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}\Big[-\frac{1}{2}(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})'\mathbf{\Sigma}_j^{-1}(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})\Big] \tag{21}$$

Setting the derivative of (21) with respect to $\mathbf{A}_{j0}$ to 0 and solving, we obtain the following:

$$\begin{aligned}
&\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}\Big[-\frac{1}{2}(\mathbf{\Sigma}_j^{-1}+\mathbf{\Sigma}_j^{-1\prime})(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})(-1)\Big]=0\\
&\Longrightarrow \sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}\Big[\frac{1}{2}(\mathbf{\Sigma}_j^{-1}+\mathbf{\Sigma}_j^{-1})(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})\Big]=0\\
&\Longrightarrow \sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}(\mathbf{\Sigma}_j^{-1})(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})=0
\end{aligned} \tag{22}$$

Analogously, setting the derivative of (21) with respect to $\mathbf{A}_{jk}$ to 0, we will get the following:

$$\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}(\mathbf{\Sigma}_j^{-1})(\mathbf{X}_b^t-\mathbf{A}_{j0}-\sum_{s=1}^{p}\mathbf{A}_{js}\mathbf{X}_b^{t-s})(\mathbf{X}_b^{t-k})'=0 \tag{23}$$

Reorganizing (22) and (23) will lead to a fully determined system, which is equivalent to the generalized Yule-Walker equations for solving the AR coefficients. The update rule for the set of AR parameters are listed below.

$$\begin{cases}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}\Big[\mathbf{A}_{j0}+\sum_{s=1}^{p}\mathbf{A}_{js}(\mathbf{X}_b^{t-s})\Big]=\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}(\mathbf{X}_b^t)\\ \sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}\Big[\mathbf{A}_{j0}(\mathbf{X}_b^{t-k})'+\sum_{s=1}^{p}\mathbf{A}_{js}(\mathbf{X}_b^{t-s})(\mathbf{X}_b^{t-k})'\Big]=\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\xi_{btj}(\mathbf{X}_b^t)(\mathbf{X}_b^{t-k})' \quad \text{for } k=1,\dots,p\end{cases}$$

3.5. Update the Bernoulli instance positive probability parameter θ

The parameter θ_j controls the probability of an instance being positive, and we denote $\delta_{btj}^l = P(I_b^t = l, Z_b^t = j|\mathbf{X}_b, Y_b, \mathbf{\Theta}')$. Collecting the terms in (9) related to θ_j, we obtain:

$$\begin{aligned}
\ell(\theta_j) &= \sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\sum_{l=0}^{1}\delta_{btj}^l\Big[l\log\theta_j+(1-l)\log(1-\theta_j)\Big]+(\zeta_1-1)\log\theta_j+(\zeta_2-1)\log(1-\theta_j)\\
&=\sum_{j=1}^{K}\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\Big[\delta_{btj}^1\log\theta_j+\delta_{btj}^0\log(1-\theta_j)\Big]+(\zeta_1-1)\log\theta_j+(\zeta_2-1)\log(1-\theta_j)\\
&=\sum_{j=1}^{K}\Big[\Big(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^1+\zeta_1-1\Big)\log\theta_j+\Big(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^0+\zeta_2-1\Big)\log(1-\theta_j)\Big]
\end{aligned} \tag{24}$$

Setting the derivative of $\ell(\theta_j)$ with respect to θ_j to 0, we will obtain the following equation

$$\begin{aligned}
&\frac{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^1+\zeta_1-1}{\theta_j}-\frac{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^0+\zeta_2-1}{1-\theta_j}=0\\
&\Longrightarrow \Big(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^1+\zeta_1-1\Big)(1-\theta_j)-\Big(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^0+\zeta_2-1\Big)\theta_j=0\\
&\Longrightarrow \Big(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^1+\zeta_1-1+\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^0+\zeta_2-1\Big)\theta_j=\sum_{b=1}^{B}\sum_{t=p+1}^{T_b}\delta_{btj}^1+\zeta_1-1
\end{aligned}$$

$$\implies \theta_j = \frac{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \delta^1_{btj} + \zeta_1 - 1}{\left(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \delta^1_{btj} + \zeta_1 - 1\right) + \left(\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \delta^0_{btj} + \zeta_2 - 1\right)} \tag{25}$$

Denote ϕ_j as shown in (26),

$$\phi_j = \frac{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \delta^1_{btj} + \zeta_1 - 1}{\sum_{b=1}^{B}\sum_{t=p+1}^{T_b} \delta^0_{btj} + \zeta_2 - 1} \tag{26}$$

Then the update rules for θ_j are shown below.

$$\theta_j = \frac{\phi_j}{1+\phi_j} \tag{27}$$

3.6. Update bag positive probability parameter ω

Collecting the terms in (9) that involve the parameter ω, we get:

$$\begin{aligned}\ell(\omega) &= \sum_{b=1}^{B}\sum_{\boldsymbol{I}} \mathbb{E}_{\mathbf{Z},\mathbf{I}|\mathbf{X},Y}\Big[\mathbb{I}(\mathbf{I}_b = \mathbf{I})\Big(Y_b \log P(Y_b = 1|\mathbf{I}_b) + (1-Y_b)\log P(Y_b = 0|\mathbf{I}_b)\Big)\Big] \\ &= \sum_{b=1}^{B}\sum_{I} P(\mathbf{I}_b = \mathbf{I}|X_{1:T_b}, Y_b)\Big(Y_b \log P(Y_b = 1|\mathbf{I}_b) + (1-Y_b)\log P(Y_b = 0|\mathbf{I}_b)\Big)\end{aligned} \tag{28}$$

As we showed in our Dynamic Programming approach, we can transform our model into an equivalent model by introducing count variables $N_b^1, \ldots, N_b^{T_b}$ to represent the counts of positive instances in bag b. Recall that $I_b^t \in \{0,1\}$. Consequently, the positive bag probability $P(Y_b = 1|\mathbf{I}_b)$ can be equivalently computed using the total number of positive instances in the bag $N_b^{T_b} = \sum_{t=1}^{T_b} I_b^t$ as follows:

$$P(Y_b = 1|N_b^{T_b}) = \frac{\sum_{t=p+1}^{T_b} I_b^t \exp(\omega I_b^t)}{\sum_{t=p+1}^{T_b} \exp(\omega I_b^t)} = \frac{N_b^{T_b}\exp(\omega)}{N_b^{T_b}\exp(\omega) + T_b - p - N_b^{T_b}} \tag{29}$$

In the denominator above, the term $T_b - p - N_b^{T_b}$ corresponds to the number of negative instances in the bag (since the instance labels are predicted starting on $t = p+1$). Using this representation and using C to represent the total number of positive instances in bag b, we can similarly rewrite (28) as:

$$\begin{aligned}\ell(\omega) = &\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b} = C|\mathbf{X}_b^{1:T_b}, Y_b)\Big[Y_b \log P(Y_b = 1|N_b^{T_b} = C) + (1-Y_b)\log P(Y_b = 0|N_b^{T_b} = C)\Big] \\ = &\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b} = C|\mathbf{X}_b^{1:T_b}, Y_b)\Big[Y_b \log\left(\frac{C\exp(\omega)}{C\exp(\omega)+T_b-p-C}\right) + (1-Y_b)\log\left(\frac{T_b-p-C}{C\exp(\omega)+T_b-p-C}\right)\Big] \\ = &\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b} = C|\mathbf{X}_b^{1:T_b}, Y_b)\Big[Y_b \log(C\exp(\omega)) - Y_b\log(C\exp(\omega)+T_b-p-C) + (1-Y_b)\log(T_b-p-C) \\ &- (1-Y_b)\log(C\exp(\omega)+T_b-p-C)\Big]\end{aligned}$$

$$
\begin{aligned}
&= \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\Big[Y_b\log(C\exp(\omega)) - Y_b\log(C\exp(\omega)+T_b-p-C) + \log(T_b-p-C) \\
&\quad - Y_b\log(T_b-p-C) - \log(C\exp(\omega)+T_b-p-C) + Y_b\log(C\exp(\omega)+T_b-p-C)\Big] \\
&= \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\Big[Y_b\log(C\exp(\omega)) + \log(T_b-p-C) - Y_b\log(T_b-p-C) \\
&\quad - \log(C\exp(\omega)+T_b-p-C)\Big] \\
&= \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\Big[Y_b\log(C\exp(\omega)) - \log(C\exp(\omega)+T_b-p-C) + \log(T_b-p-C) \\
&\quad - Y_b\log(T_b-p-C)\Big]
\end{aligned}
$$

Ignoring the terms that don't involve ω, we get:

$$
\begin{aligned}
&\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\Big[Y_b\cdot\omega - \log\left(C\exp(\omega)+T_b-p-C\right)\Big] \\
&= \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)Y_b\cdot\omega - \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\log\left(C\exp(\omega)+T_b-p-C\right) \\
&= \sum_{b=1}^{B} Y_b\cdot\omega - \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\log\left(C\exp(\omega)+T_b-p-C\right) \qquad (30)
\end{aligned}
$$

The first-order gradient of the (30) with respect to ω is:

$$
\begin{aligned}
l'(\omega) &= \sum_{b=1}^{B} Y_b - \sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\frac{C\exp(\omega)}{C\exp(\omega)+T_b-p-C} \\
&= \sum_{b=1}^{B}\Big[Y_b - \sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)P(Y_b=1|N_b^{T_b}=C)\Big] \qquad (31)
\end{aligned}
$$

The second-order gradient of the (30) with respect to ω is:

$$
\begin{aligned}
l''(\omega) &= -\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\frac{C\exp(\omega)(C\exp(\omega)+T_b-p-C) - C\exp(\omega)C\exp(\omega)}{(C\exp(\omega)+T_b-p-C)^2} \\
&= -\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\frac{C\exp(\omega)[C\exp(\omega)+T_b-p-C-C\exp(\omega)]}{(C\exp(\omega)+T_b-p-C)^2} \\
&= -\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\frac{C\exp(\omega)(T_b-p-C)}{(C\exp(\omega)+T_b-p-C)^2} \\
&= -\sum_{b=1}^{B}\sum_{C=0}^{T_b-p} P(N_b^{T_b}=C|\mathbf{X}_b^{1:T_b},Y_b)\Big[P(Y_b=1|N_b^{T_b}=C)\Big(1-P(Y_b=1|N_b^{T_b}=C)\Big)\Big] \qquad (32)
\end{aligned}
$$

Combining Equation 31 and 32, the update rule for ω with a newton step is $\omega_{new} = \omega_{old} - \dfrac{l'(\omega_{old})}{l''(\omega_{old})}$. For the efficiency of M-step update, we choose to update the ω with only one gradient step per E-M iteration.

4. Experiment Results

4.1. Running Time Illustration

We demonstrate the running time comparison with/without the dynamic programing speedup with an illustrative example.

The exhaustive enumeration in the E-step has a runtime complexity of $O(B(2K)^T)$, where K is the number of clusters and T is the bag length. In contrast, the dynamic programming approach we proposed runs in polynomial time $O(BK^2T^2)$. On Fig. 4 on the left, to keep the running time of exhaustive enumeration under 24 hours, we varied the bag length from 4 to 10 with a total of 20 bags.

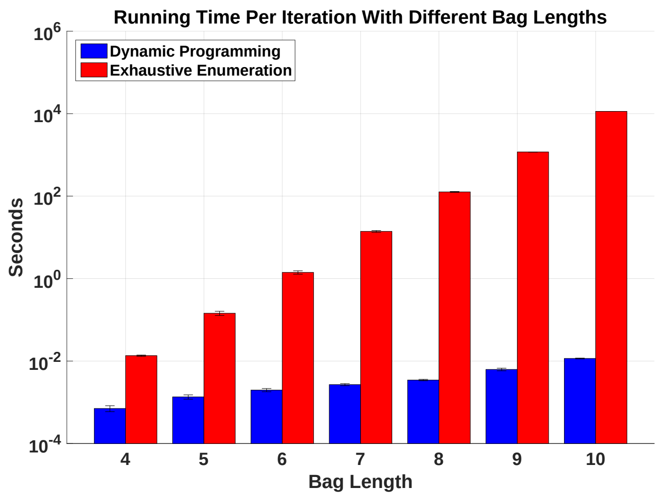

Figure 4. Left: The running time comparison between the exhaustive approach and dynamic programming. The black error bars denote the 95% confidence interval.

The experiment at each bag length is repeated 10 times, and the average running time per E-M iteration is reported. We make the scale on the y-axis to grow exponentially on Fig. 4, so it is clear to see that the dynamic programming approach we proposed has made the exact inference by running in quadratic time; by contrast, the exhaustive enumeration is running in exponential time.