INFERENCE AND IMPUTATION FOR HIDDEN MARKOV MODELS WITH HYBRID STATE OUTPUTS

ANNA HAENSCH

Abstract. Hidden Markov models are an effective tool for classifying sequential data and can be trained on multiple episodic data types. In this note we derive the equations necessary to carry out parameter optimization for hidden Markov models using expectation maximization. We do this in a flexible framework, allowing for mutiple observation types coming from discrete and Gaussian mixture model distributions, and multiple disjoint observation episodes. We will also present several methods for imputation using HMMs.

1. Introduction

One important task in machine learning is to detect latent signals present in data and classify data accordingly. In the case of timeseries data, classifications should capture not only clustering across spatial (in the generalized metric sense) dimensions but sequential dynamics as well. The hidden Markov model (HMM) is a simple yet highly effective model architecture for classifying data according to both spatial and temporal dynamics. In an HMM, latent states are drawn according to a joint probability distribution depending on the observed data at time, t, and a Markov transition probability.

Let's consider for example, the 2-dimensional cluster shown in Figure 1.

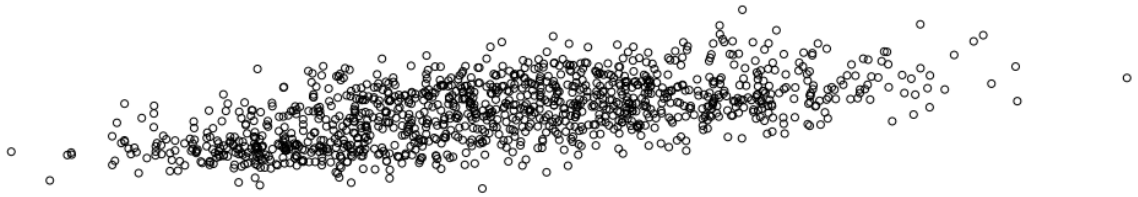
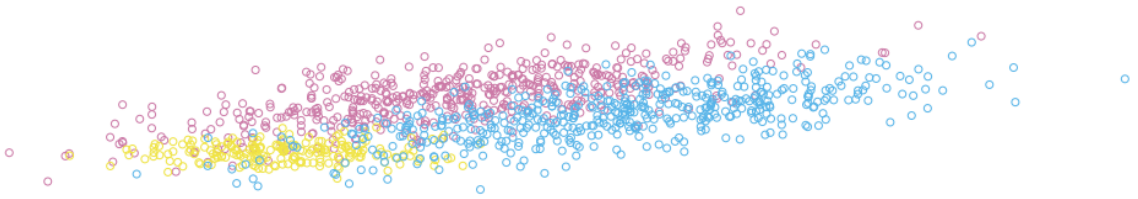


Figure 1

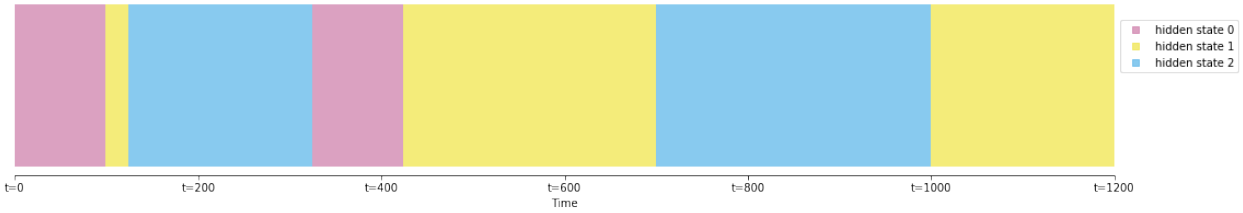
From a purely spatial perspective, this data appears to be drawn from a Gaussian distribution with high covariance. However, if we consider similarly arranged data with an added time component, several obvious clusters emerge as in Figure 2. The goal of hidden Markov modeling is to detect these underlying time dependent states and learn not only the spatial clustering parameters, but also the transition parameters. We refer to the unlabeled data in Figure 1 as *observations*, and the sequence of latent cluster assignments shown in Figure 2b as the *hidden states*.

Key words and phrases. Hidden Markov model, statistical inference.

The author also thanks her colleagues at Tagup Inc. in Somerville, MA for helpful discussions during her time there in the 2019-2020 academic year. *Uploaded: August 9, 2021.*



(A) Spatial clusters by hidden state for observations from time $t = 0$ to $t = 1200$.



(B) Temporal dynamics of hidden state transitions from time $t = 0$ to $t = 1200$.

FIGURE 2

For a fixed model class, one goal of modeling is to find a set of model parameters, θ, that maxmize the likelihood function,

$$L(\theta; X) = p(X \mid \theta), \tag{1}$$

for a set of observed data, X. In the case of HMMs, we also have a sequence of hidden states, Z. Marginalizing over the hidden states, we can write the likehood as

$$L(\theta; X) = \int p(X, Z \mid \theta)\, dZ \tag{2}$$

where the integrand above is called the *complete data likelihood*.

In an HMM, the sequence of latent states, Z, must satisfy the Markov property, that is, the latent state at any time t is dependent on the latent state immediately prior, but is independent of the latent state evolution prior to time $t - 1$. This conditional dependence can be viewed as a directed graph as in Figure 3, where Z_t denotes the hidden state and the X_t denotes the observed data at time t.

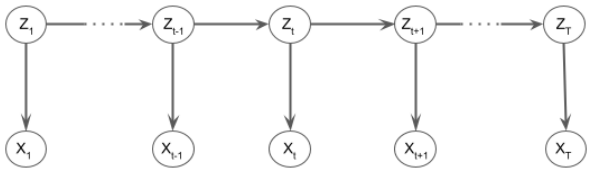


FIGURE 3. HMM graphical structure.

2. Preliminaries and Notation

In what follows, we will refer the random variables as upper case Latin letters such as X, precise value as lower case Latin letters such as x, and the probability distributions from which they are drawn as caligraphic letters such as $\mathcal{X}$.

The observation, X_t, at time t can be viewed as a random variable drawn from the distribution $\mathcal{X}$, which may consist of mutliple statistically independent probability distributions. Suppose $\mathcal{X}$ is a product of K distributions,

$$\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_K,$$

then each X_t has K components,

$$X_t = (X_{t,1}, ..., X_{t,K})$$

where

$$X_{t,k} \sim \mathcal{X}_k.$$

We will denote a sequence of T random variables drawn from $\mathcal{X}$ by

$$X_{1:T} := \{X_1, ..., X_T\}.$$

Recall that the goal of hidden Markov modeling is to find the mostly likely sequence of latent states, $Z_{1:T}$, corresponding to a sequence of observations, $X_{1:T}$. We will focus our attention on the case where Z_t is drawn from a disrcete distribution, $\mathcal{Z}$, containing N elements

$$\{h_1, ..., h_N\}.$$

Of course it is possible, and of considerable interest, to choose more complicated latent state distributions, but we will remain in the most simple case; for further details, the reader is directed to Chapter 13 of Bishop's influential text [1]. The sequence of latent states must satisfy the Markov property, which we can state concretely as

$$p(Z_t \mid Z_{t-1}) \quad = \quad p(Z_t \mid Z_1, ..., Z_{t-1}). \tag{3}$$

This so-called *transition probability* is one of several parameters of the HMM. The transition probability is given by a $N \times N$ right stochastic transition matrix, A, where the ij^{th} entry of A is

$$A_{ij} \quad = \quad p(Z_t = h_j \mid Z_{t-1} = h_i).$$

for $1 < t \leq T$, satisfying

$$\sum_{j=1}^{N} A_{ij} \quad = \quad 1. \tag{4}$$

When $t = 1$, the *initial state probability* is given by the $1 \times N$ array, π, where the i^{th} entry of π is

$$\pi_i \quad = \quad p(X_1 = h_i)$$

for $1 \leq i \leq N$ satisfying

$$\sum_{i=1}^{N} \pi_i \quad = \quad 1. \tag{5}$$

Finally, the conditional probability joining the latent states and observed data is called the *emission probability* and it is given by

$$p(X_t = x_t \mid Z_t = z_t) \quad = \quad \prod_{k=1}^{K} p(X_{t,k} = x_{t,k} \mid Z_t = z_t).$$

Here the probability for term k depends on the choice of distribution for $\mathcal{X}_k$ so the emission probability is given by K distinct model parameters, $\theta_1, ..., \theta_K$. In sections 3.3.1 and 3.3.2 we give a more concrete form to these θ_k for discrete, gaussian mixture model, and eventually hybrid state emissions.

An HMM is fully parameterized as

$$\theta = \{\pi, A, \theta_1, ..., \theta_K\}.$$

Given a set of model parameters, we can write the probabilities above in a parametric form,

$$\begin{aligned} p_0(z_1; \pi) &:= p(Z_1 = z_1) \\ p_t(z_t \mid z_{t-1}; A) &:= p(Z_t = z_t \mid Z_{t-1} = z_{t-1}) \\ p_k(x_t \mid z_t; \theta_k) &:= p(X_{t,k} = x_{t,k} \mid Z_t = z_t) \end{aligned}$$

for $1 \leq k \leq K$. The parameters will often be omitted when it is clear from context.

3. Learning optimal model parameters

Let θ denote a fully parameterized HMM. From equation (2) we know that the likelihood of θ is maximized by finding

$$\theta^* = \underset{\theta}{\operatorname{argmax}} \int p(X_{1:T}, Z_{1:T} \mid \theta)\ dZ_{1:T}.$$

Unfortunately, this integral is often quite difficult to compute in practice. Even in the cases where it is possibly to compute analytically, it can still be computationally intractable. Moreover, solving the maximization problem also involves optimizing the quantity on the right-hand side, which can often involve complicated non-convex optimization. In what follows we will see how a the Markov property, coupled with dynamic programming allows us to bypass this complicated problem.

Rather than compute the marginal probabilities directly, we will make use of the Baum-Welch variant of the classical *expectation-maximization* (EM) algorithm to find an optimal set of model parameters. The EM algorithm consists of an expectation step (or E-step) and a maximization (or M-step). For the E-step we define an auxilary function, Q, of current parameters, θ' and new parameters, θ,

$$Q(\theta, \theta') = \mathbb{E}_{Z_{1:T}|X_{1:T},\theta'} \left[\log p(X_{1:T}, Z_{1:T} \mid \theta)\right], \tag{6}$$

which gives the expected value of the complete data log likelihood with respect to the sequence of hidden states $Z_{1:T}$ given $X_{1:T}$ and the current model parameters. For the M-step, instead of maximizing the likelihood function (2) directly, we iteratively optimize Q by computing

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta').$$

In what follows, we show that iteratively maximizing this auxiliary function is sufficient to find a local maximum for the likelihood function. Since it will simplify the discussion significantly, we will work in terms of the log likelihood function,

$$\mathcal{L}(\theta; X_{1:T}) := \log p(X_{1:T} \mid \theta) = \log \int p(X_{1:T}, Z_{1:T} \mid \theta)\ dZ_{1:T}.$$

Theorem 3.1. *If $Z_{1:T}$ is drawn from a discrete probability distribution, then*

$$Q(\theta, \theta') \leq \mathcal{L}(\theta, X_{1:T}) \tag{7}$$

for any fixed set of model parameters θ'. Moreover,

$$Q(\theta, \theta') \geq Q(\theta', \theta') \tag{8}$$

implies

$$\mathcal{L}(\theta, X_{1:T}) \geq \mathcal{L}(\theta', X_{1:T}). \tag{9}$$

Proof. Suppose that $Z_{1:T}$ is drawn from a discrete probability distribution. From Bayes' rule, we know that

$$\log p(X_{1:T} \mid \theta) = \log p(X_{1:T}, Z_{1:T} \mid \theta) - \log p(Z_{1:T} \mid X_{1:T}, \theta).$$

Multiplying both sides of this equation by $p(Z_{1:T} \mid X_{1:T}, \theta')$ and integrating over the space of all possible $Z_{1:T}$, we get

$$\mathcal{L}(\theta; X_{1:T}) = Q(\theta, \theta') - \int \log p(Z_{1:T} \mid X_{1:T}, \theta) \cdot p(Z_{1:T} \mid X_{1:T}, \theta') \; dZ_{1:T}.$$

The negative term on the right hand side above is just the entropy of the random variables X and Z, which in the case of discrete Z is always positive, from which we obtain (7).

The equation above still holds if we replace θ with θ', in particular,

$$\mathcal{L}(\theta'; X_{1:T}) = Q(\theta', \theta') - \int \log p(Z_{1:T} \mid X_{1:T}, \theta') \cdot p(Z_{1:T} \mid X_{1:T}, \theta') \; dZ_{1:T}.$$

Subtracting these equations from one another we obtain

$$\begin{aligned} \mathcal{L}(\theta; X_{1:T}) - \mathcal{L}(\theta'; X_{1:T}) &= Q(\theta, \theta') - Q(\theta', \theta') + ... \\ &\quad ... + \int \log \frac{p(Z_{1:T} \mid X_{1:T}, \theta')}{p(Z_{1:T} \mid X_{1:T}, \theta)} \cdot p(Z_{1:T} \mid X_{1:T}, \theta') \; dZ_{1:T}. \end{aligned}$$

However, we recognize the second term on the right hand side above as the Kullback-Liebler divergence,

$$D_{KL}(p(Z_{1:T} \mid X_{1:T}, \theta') \mid\mid p(Z_{1:T} \mid X_{1:T}, \theta),$$

which is always greater than or equal to 0. In particular, this shows that if Q increases, then $\mathcal{L}$ increases at least as much. □

Corollary 3.2. *If $Z_{1:T}$ is drawn from a discrete probabilty distribution, then repeated iterative improvements of Q will enventually lead to a local maximum for $\mathcal{L}$.*

Proof. Since Q is bounded above by 0 and since the input space of Q is compact, iteratively improving Q will eventually lead to a critical point for Q. That is, we will reach a set of model parameters, θ^*, such that

$$Q(\theta, \theta^*) \leq Q(\theta^*, \theta^*)$$

for any choice of θ. If we consider $Q(\theta, \theta^*)$ as a function in θ, then this is equivalent to

$$\frac{d}{d\theta} \left[Q(\theta, \theta^*) \right] \Big|_{\theta^*} = 0. \tag{10}$$

On the other hand, taking the dervative of $\mathcal{L}$ with respect to θ and evaluating at θ^* gives

$$\frac{d}{d\theta}\mathcal{L}(\theta; X_{1:T}) = \frac{1}{L(\theta^*; X_{1:T})} \sum_{Z_{1:T}} \frac{d}{d\theta} \left[p(Z_{1:T}, X_{1:T} \mid \theta)\right]\Big|_{\theta^*} \tag{11}$$

Now, combining equations (10) and (11) above, we obtain

$$\begin{aligned}
\frac{d}{d\theta}\left[Q(\theta, \theta^*)\right]\Big|_{\theta^*} &= \frac{d}{d\theta}\left[\int \log p(X_{1:T}, Z_{1:T} \mid \theta) p(Z_{1:T} \mid X_{1:T}, \theta^*)\, dZ_{1:T}\right]\Big|_{\theta^*} \\
&= \sum_{Z_{1:T}} p(Z_{1:T} \mid X_{1:T}, \theta^*) \cdot \frac{d}{d\theta}\left[\log p(Z_{1:T}, X_{1:T} \mid \theta)\right]\Big|_{\theta^*} \\
&= \sum_{Z_{1:T}} \frac{p(Z_{1:T} \mid X_{1:T}, \theta^*)}{p(Z_{1:T}, X_{1:T} \mid \theta^*)} \cdot \frac{d}{d\theta}\left[p(Z_{1:T}, X_{1:T} \mid \theta)\right]\Big|_{\theta^*} \\
&= \frac{1}{p(X_{1:T} \mid \theta^*)} \cdot \sum_{Z_{1:T}} \frac{d}{d\theta}\left[p(Z_{1:T}, X_{1:T} \mid \theta)\right]\Big|_{\theta^*} \\
&= \frac{1}{L(\theta^*; X_{1:T})} \cdot \frac{d}{d\theta}\left[L(\theta; X_{1:T})\right]\Big|_{\theta^*} \\
&= \frac{d}{d\theta}\left[\mathcal{L}(\theta; X_{1:T})\right]\Big|_{\theta^*},
\end{aligned}$$

from which it follows that θ^* is a critical point for $\mathcal{L}$. Therefore, we have shown that iteratively improving Q will eventually lead to a local maximum, θ^*, which will also be a critical point for $\mathcal{L}$. Moreover, from Lemma 3.1 we know that this critical point will be a local maximum. □

With successive iterations of the E- and M-steps, the sequence

$$\{Q(\theta^i, \theta^{i-1}) : i \in \mathbb{Z}^+\}$$

is bounded and monotone increasing. Since it is bounded and therefore guaranteed to converge, we will eventually reach a local maximum for L by Corollary 3.2. However, it still remains to be shown how to compute the expected value in the E-step and how to carry out the maximization described in the M-step. In the following sections we will carefully derive all of the quantities necessary to carry out these steps, but a reader wishing to skip directly to the punchline is directed to the pseudocode for the flexible EM algorithm with hybrid emissions, Algorithm 2.

3.1. **The forward and backward algorithms.** From the graphical model in Figure 3, it can be verified using the *d-separation* algorithm that $X_{1:t}$ and $X_{t+1:T}$ are conditionaly independent given Z_t. Thefore, at any time t, we have

$$p(Z_t, X_{1:T} \mid \theta) \;=\; p(Z_t, X_{1:t} \mid \theta) \cdot p(X_{t+1,T} \mid Z_t, \theta)$$

allowing for the use of dynamic programming to avoid direct computation of the joint probability necessary for the E-step. For a fixed set of observations, $x_{1:T}$, we will define

$$\alpha_t(z_t) \;:=\; p(Z_t = z_T, X_{1:t} = x_{1:t} \mid \theta)$$

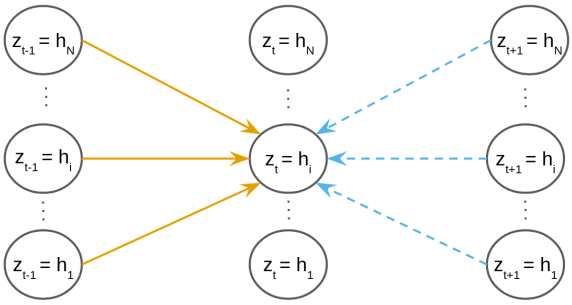


FIGURE 4. This schematic shows a single step of the forward-backward algorithm in terms of the forward pass as solid orange lines and the backward pass as dashed blue lines.

and

$$\beta_t(z_t) := p(X_{t+1:T} = x_{t+1:T} \mid Z_t = z_t, \theta),$$

whereby

$$p(Z_t = z_t, X_{1:T} = x_{1:T} \mid \theta) = \alpha_t(z_t) \cdot \beta_t(z_t). \tag{12}$$

Using the model paramaterization, θ, we also know that the full joint probability can be expressed as

$$p(x_{1:T}, z_{1:T} \mid \theta) = p_o(z_1) \cdot \prod_{t=2}^{T} p_t(z_t \mid z_{t-1}) \cdot \prod_{t=1}^{T} \prod_{k=1}^{K} p_k(x_t \mid z_t) \tag{13}$$

for a given sequence of observations, $x_{1:T}$, and hidden states, $z_{1:T}$. Using this, we can initialize α with

$$\alpha_1(z_1) = p_0(z_1) \cdot \prod_{k=1}^{K} p_k(x_1 \mid z_1)$$

and define remaining terms recursively as

$$\alpha_t(z_t) = \prod_{k=1}^{K} p_k(x_t \mid z_t) \int p_t(z_t \mid z_{t-1}) \cdot \alpha_{t-1}(z_{t-1}) \, dz_{t-1}$$

for $1 < t \leq T$. Similary, we initialize β with

$$\beta_T(z) = 1$$

and define remaining terms recursively as

$$\beta_t(z_t) = \int \prod_{k=1}^{K} p_k(x_{t+1} \mid z_{t+1}) \cdot p_t(z_{t+1} \mid z_t) \cdot \beta_{t+1}(z_{t+1}) \, dz_{t+1}$$

for $1 \leq t < T$, taken in descending order. These recursively defined values for α and β are the core of the *forward-backward algorithm*, given here as Algorithm 1. The forward backward algorithm considers the possibility of each hidden state, h_i, at each time, t, by computing the cumulative likelihood of arriving in state h_i given any of the possible prior states (this is the forward part, seen as solid orange lines in Figure 4) and any of the possible subsequent states (this is the backward part, seen as dashed blue lines in Figure 4).

Algorithm 1 Discrete Hidden State Forward-Backward Algorithm

Input Model, θ; observations, $x_{1:T}$.
Output Forward and backward probability arrays, α and β.

procedure ForwardBackward($\theta, x_{1:T}$)
  Initialize empty $(T \times N)$-dimensional arrays α and β.
  $\alpha[0, i] \leftarrow \pi_i \cdot \prod_{k=1}^{K} p_k(x_{k,1} \mid h_i)$ for $1 \leq i \leq N$
  for all timesteps $t \in \{2, ..., T\}$ **do**
    for all hidden states $i \in \{1, ..., N\}$ **do**
      $\alpha[t, i] \leftarrow \prod_{k=1}^{K} p_k(x_{k,t} \mid h_i) \cdot \sum_{j=1}^{N} A_{ji} \cdot \alpha[t-1, j]$
    end for
  end for
  $\beta[T, i] \leftarrow 1$ for $1 \leq i \leq N$.
  for all timesteps $t \in \{T-1, ..., 1\}$ **do**
    for all hidden states $i \in \{1, ..., N\}$ **do**
      $\beta[t, i] \leftarrow \sum_{j=1}^{N} \prod_{k=1}^{K} p_k(x_{k,t+1} \mid j) \cdot A_{ij} \cdot \beta[t+1, j]$
    end for
  end for
  return α, β
end procedure

With values for α and β in hand, using equation (12) and marginalizing over the hidden state space, we obtain

$$p(Z_t = z \mid x_{1:T}, \theta) = \frac{p(z, x_{1:T} \mid \theta)}{\int p(z', x_{1:T} \mid \theta)\ dz'} = \frac{\alpha_t(z) \cdot \beta_t(z)}{\int \alpha_t(z') \cdot \beta_t(z')\ dz'}.$$

Since we are only concerned with hidden states drawn from a discrete state space, we will define the following notation to express this more concisely. We define the $N \times 1$ vector $\langle Z_t \rangle$, whose i^{th} entry is $p(Z_t = h_i \mid x_{1:T}, \theta)$, which can be written in terms of α and β as

$$\langle Z_t \rangle_i \quad = \quad \frac{\alpha_t(h_i) \cdot \beta_t(h_i)}{\sum_{j=1}^{N} \alpha_t(h_j) \cdot \beta_t(h_j)}. \tag{14}$$

We will define the $N \times N$ vector $\langle Z_{t-1} Z_t \rangle$ whose ij^{th} entry is $p(Z_{t-1} = h_i, Z_t = h_j \mid x_{1:T}, \theta)$, which can be written in terms of α and β as

$$\langle Z_{t-1} Z_t \rangle_{ij} \quad = \quad \frac{\alpha_{t-1}(h_i) \cdot A_{ij} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j) \cdot \beta_t(h_j)}{\sum_{j'=1}^{N} \alpha_{t-1}(h_i) \cdot A_{ij'} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_{j'}) \cdot \beta_t(h_{j'})}. \tag{15}$$

These quantities in equations (14) and (15) are often referred to as γ and ξ, respectively, but we will be adopting the notation of [2].

3.2. **Computing the expected value.** Using the values for $\langle Z_t \rangle$ and $\langle Z_{t-1} Z_t \rangle$ computed in the previous section, we now have a more concise way to express the expected value in equation (6). First, we observe

$$\mathbb{E}_{Z_{1:T} | X_{1:T}, \theta'} \left[\log p(X_{1:T}, Z_{1:T} \mid \theta) \right] \quad = \quad \int \log p(X_{1:T}, Z_{1:T} \mid \theta) \cdot p(Z_{1:T} \mid X_{1:T}, \theta')\ dZ_{1:T}.$$

From (13), we know

$$\log p(X_{1:T}, Z_{1:T} \mid \theta) = \log p(Z_1 \mid \theta) + \sum_{t=2}^{T} \log p(Z_t \mid Z_{t-1}, \theta) + \sum_{t=1}^{T} \log p(X_t \mid Z_t, \theta).$$

In this way the expected value naturally decomposes into three linearly independent terms, which we will deal with one at a time. Starting with the initial state term, we obtain

$$\int \log p(Z_1 \mid \theta) \cdot p(Z_{1:T} \mid X_{1:T}, \theta')\ dZ_{1:T} = \sum_{i=1}^{N} \log p_0(h_i) \cdot \langle Z_1 \rangle_i = \sum_{i=1}^{N} \log \pi_i \cdot \langle Z_1 \rangle_i.$$

For the transition term, we obtain

$$\begin{aligned} & \int \sum_{t=2}^{T} \log p(Z_t \mid Z_{t-1}, \theta) \cdot p(Z_{1:T} \mid X_{1:T}, \theta')\ dZ_{1:T} \\ = & \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \log p_t(h_j \mid h_i) \cdot \langle Z_{t-1} Z_t \rangle_{ij} \\ = & \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \log A_{ij} \cdot \langle Z_{t-1} Z_t \rangle_{ij}, \end{aligned}$$

and for the emission term,

$$\begin{aligned} & \int \sum_{t=1}^{T} \log p(X_t \mid Z_t, \theta) \cdot p(Z_{1:T} \mid X_{1:T}, \theta')\ dZ_{1:T} \\ = & \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{K} \log p_k(x_t \mid h_i) \cdot \langle Z_t \rangle_i. \end{aligned}$$

We will note that the $\langle \cdot \rangle$ terms here are taken with respect to the θ' model parameters. Combining all of the terms above, the expected value can be efficiently computed as

$$\begin{aligned} \mathbb{E}_{Z_{1:T} \mid X_{1:T}, \theta'} \left[\log p(X_{1:T}, Z_{1:T} \mid \theta) \right] = & \sum_{i=1}^{N} \log \pi_i \cdot \langle Z_t \rangle_i + ... \\ & ... + \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \log A_{ij} \cdot \langle Z_{t-1} Z_t \rangle_{ij} + ... \\ & ... + \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{k=1}^{K} \log p_k(x_t \mid h_i) \cdot \langle Z_t \rangle_i. \end{aligned}$$

Now we are well positioned to maximize this simplified expression.

3.3. **Maximizing the expected value.** For a complete current set of model parameters, θ', the goal is to find an optimal set of model parameters

$$\theta^* = \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{Z_{1:T} \mid X_{1:T}, \theta'} \left[\log p(X_{1:T}, Z_{1:T} \mid \theta) \right] \tag{16}$$

thereby maximizing the expected value. Once obtained, the current model parameters are updated, and one iteration of EM is complete. In the previous section we obtained a simplified expression for the expected value, separated into three components describing the contribution of the initial state, the transition, and the emission probabilities to the conditional expected value. Because of the linearity of the expected value, each of these terms in the expected value can be treated as a separate optimization problem.

In some cases certain constraints must be satisfied, for example the initial state probability is subject to equation (5) and the transition probability is subject to equation (4). In some cases the emission probability will also realize constraints, but but these will depend largely on the distribution at hand and will be discussed more fully in sections 3.3.1 and 3.3.2.

Now it only remains to derive precise formulations for the so-called *update equations* for the model parameters which we will deal with one by one, beginning with the initial state parameter. From the previous section we know that the contribution to the expected value from the initial state probaility is maximized for π^* satisfying

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \sum_{i=1}^{N} \log \pi_i \cdot \langle Z_1 \rangle_i.$$

To find this value, we will compute the gradient of

$$\sum_{i=1}^{N} \log \pi_i \cdot \langle Z_1 \rangle_i - \lambda \left(-1 + \sum_{i=1}^{N} \pi_i \right)$$

where λ is a Lagrange multipier, and solve. Computing the gradient and setting it equal to 0 we obtain

$$\frac{1}{\pi_i} \cdot \langle Z_1 \rangle_i - \lambda = 0 \qquad \text{and} \qquad 1 - \sum_{i=1}^{N} \pi_i = 0.$$

Combining the equations above, we arrive at

$$\pi_i = \frac{\langle Z_1 \rangle_i}{\sum_{j=1}^{N} \langle Z_1 \rangle_j} = \langle Z_1 \rangle_i,$$

and therefore the optimal initial state parameter is given by

$$\pi_i^* = \langle Z_1 \rangle_i \tag{17}$$

for $1 \leq i \leq N$.

To update the transition probability, we solve for A^* satisfying

$$A^* = \underset{A}{\operatorname{argmax}} \sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \log A_{ij} \cdot \langle Z_{t-1} Z_t \rangle_{ij}.$$

To find this value, we will proceed precisely as in the previous paragraph, computing the gradient of

$$\sum_{t=2}^{T} \sum_{i=1}^{N} \sum_{j=1}^{N} \log A_{ij} \cdot \langle Z_{t-1} Z_t \rangle_{ij} - \lambda \left(-1 + \sum_{j=1}^{N} A_{ij} \right)$$

and solving. Computing the gradient and setting it equal to 0 we obtain

$$\frac{1}{A_{ij}}\sum_{t=2}^{T}\langle Z_{t-1}Z_t\rangle_{ij} - \lambda = 0 \qquad \text{and} \qquad 1 - \sum_{j=1}^{N} A_{ij} = 0.$$

The equation on the left can be rearranged to get

$$A_{ij} \;=\; \frac{1}{\lambda}\sum_{t=2}^{T}\langle Z_{t-1}Z_t\rangle_{ij},$$

and combined with the equation on the right for

$$1 = \sum_{j=1}^{N} A_{ij} = \frac{1}{\lambda}\sum_{j=1}^{N}\sum_{t=2}^{T}\langle Z_{t-1}Z_t\rangle_{ij}.$$

From here we get

$$\lambda \;=\; \sum_{t=2}^{T}\sum_{j=1}^{N}\langle Z_{t-1}Z_t\rangle_{ij}.$$

which can be solved for to arrive at the optimal transition parameter

$$A^*_{ij} \;=\; \frac{\sum_{t=2}^{T}\langle Z_{t-1}Z_t\rangle_{ij}}{\sum_{t=2}^{T}\sum_{j'=1}^{N}\langle Z_{t-1}Z_t\rangle_{ij'}} \tag{18}$$

for $1 \le i, j \le N$.

Since it will require the introduction of some new notation, we will consider the emission probabilities in separate sections. In section 3.3.1 we will deal with discrete observations and in section 3.3.2 we will deal with observations drawn from a Gaussian mixture model. Finally, in section 3.4 we will combine all of these notions for a maximally flexible set of update equations that include multiple observation types across several disjoint episodes of observation.

3.3.1. *Discrete emissions.* Suppose we have K discrete emission features, and suppose feature k takes on one of D_k values, for $1 \le k \le K$. Then there are

$$D = \prod_{k=1}^{K} D_k$$

possible discrete observable vectors. The conditional probability of observing vector

$$d = (d_1, ..., d_K)$$

at time, t, given the hidden state h_i is express concisely as a $D \times N$ array, B, where the di^{th} entry is

$$B_{di} = p(X_t = d \mid Z_t = h_i, \theta) = \prod_{k=1}^{K} p(X_{t,k} = d_k \mid Z_t = h_i, \theta)$$

and consequently

$$\sum_{d=1}^{D} B_{di} = 1. \tag{19}$$

In this case, the contribution of the emission probabilty to the expected value can be written as

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{k=1}^{K}\log p_k(x_t \mid h_i)\cdot\langle Z_t\rangle_i = \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{d=1}^{D}\delta_{d,X_t}\log B_{di}\cdot\langle Z_t\rangle_i.$$

where δ denotes the Kronecker delta function,

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Therefore to optimize this portion of the expected value, our goal is to find an optimal parameter, B^*, such that

$$B^* = \underset{B}{\operatorname{argmax}} \sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{d=1}^{D}\delta_{d,X_t}\log B_{di}\cdot\langle Z_t\rangle_i.$$

In this case, we are optimizing subject to constraints of equation (19), therefore we once again proceed by computing the gradient of

$$\sum_{t=1}^{T}\sum_{d=1}^{D}\sum_{i=1}^{N}\delta_{d,X_t}\cdot\log B_{di}\cdot\langle Z_t\rangle_i - \lambda\left(-1+\sum_{d=1}^{D}B_{di}\right),$$

where λ is the Lagrange multiplier, and solving. Computing the gradient and setting it equal to 0, we obtain

$$\frac{1}{B_{di}}\sum_{t=1}^{T}\delta_{d,X_t}\cdot\langle Z_t\rangle_i - \lambda = 0 \qquad \text{and} \qquad 1-\sum_{d=1}^{D}B_{di} = 0.$$

The equation on the left can be rearranged to get

$$B_{di} = \frac{1}{\lambda}\sum_{t=1}^{T}\delta_{d,X_t}\cdot\langle Z_t\rangle_i$$

and combined with the equation on the right, for

$$1 = \sum_{d=1}^{D}B_{di} = \frac{1}{\lambda}\sum_{d=1}^{D}\sum_{t=1}^{T}\delta_{d,X_t}\cdot\langle Z_t\rangle_i.$$

Solving this equation for λ and substituting it into the prior equation, we can solve to obtain the optimal emission parameter

$$B^*_{di} = \frac{\sum_{t=1}^{T}\delta_{d,X_t}\cdot\langle Z_t\rangle_i}{\sum_{d'=1}^{D}\sum_{t=1}^{T}\delta_{d',X_t}\cdot\langle Z_t\rangle_i} \tag{20}$$

for $1 \le d \le D$ and $1 \le i \le N$.

3.3.2. *Gaussian mixture model emissions.* Suppose the observations are drawn from an M component Gausssian mixture model. For a K-dimensional multivariate Gaussian, the probability density function is given by

$$\mathcal{N}(x_t; \mu, \Sigma) = \frac{\exp\left\{-\frac{1}{2}(x_t - \mu)^\intercal \cdot \Sigma^{-1} \cdot (x_t - \mu)\right\}}{\sqrt{(2\pi)^K \mid \Sigma \mid}}$$

where μ and Σ are means and coviariance matrices, repsectively, and $^\intercal$ denotes the matrix transpose. The Gaussian mixture model consists of M components with indpendent probabilty distributions with an attached component weight. Conditioning on a hidden state, h_i, the component weight, W_{im}, is the probability of component m given hidden state h_i, from which it follows that

$$\sum_{m=1}^{M} W_{im} = 1, \tag{21}$$

where W is the $N \times M$ array of all weights. Therefore, the probability of an observation conditioned on hidden state h_i at time t, is given by

$$p(X_t = x_t \mid Z_t = h_i) = \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im})$$

where μ_{im} and Σ_{im} are means and covariance matrices, respectively, for component m given hidden state h_i. Let μ and Σ denote the set of all means and covariances, respectively. Therefore, in this setting, the full set of HMM model parameters is given by

$$\theta = \{A, \pi, W, \mu, \Sigma\}.$$

We already learned how to optimize A and π in section 3.3. Now we just need to determine the proper update equations for the remaining terms, which will be given by

$$W^*, \mu^*, \Sigma^* = \underset{W,\mu,\Sigma}{\operatorname{argmax}} \sum_{t=1}^{T} \sum_{i=1}^{N} \log \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i$$

Because it will simplify the exposition in what follows, we introduce one final piece of bracket notation, $\langle Z_t X_t \rangle$, which denotes the $N \times M$ array, where the im^{th} entry,

$$\langle Z_t X_t \rangle_{im} = \frac{W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i}{\sum_{m'=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})}, \tag{22}$$

is the probability of hidden state h_i and Gaussian mixture component m generating the observation at time t.

To determine the optimal weight parameters, will optimize the contribution of the emission to the expected value contrained by (21), by computing the gradient of

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \log \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i - \lambda \left(-1 + \sum_{m=1}^{M} W_{im} \right),$$

with respect to each of the W_{im} and the Lagrange multiplier, λ, and solving. Computing the gradient and setting each of the terms equal to 0, we get

$$\sum_{t=1}^{T} \frac{\mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i}{\sum_{m'=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})} - \lambda = 0 \text{ and } 1 - \sum_{m=1}^{M} W_{im} = 0.$$

Rearranging the equation on the left we get

$$W_{im} = \frac{1}{\lambda} \sum_{t=1}^{T} \frac{W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i}{\sum_{m'=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})}$$

which is combined with equation on the right to get

$$\lambda = \lambda \sum_{m=1}^{M} W_{im} = \sum_{t=1}^{T} \langle Z_t \rangle_i.$$

From here we solve for W_{im} to get the optimal weight parameter

$$W_{im}^* = \frac{\sum_{t=1}^{T} \langle Z_t X_t \rangle_{im}}{\sum_{t=1}^{T} \langle Z_t \rangle_i} \tag{23}$$

for $1 \leq i \leq N$ and $1 \leq m \leq M$.

To find the optimal mean parameters, we need to compute the derivative of

$$\sum_{t=1}^{T} \sum_{i=1}^{N} \log \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot \langle Z_t \rangle_i, \tag{24}$$

with respect to μ_{im}, and solve. Differentiating with respect to W_{im}, setting the result equal to 0 and dividing by unnecessary constant terms, we obtain

$$0 = \sum_{t=1}^{T} \frac{W_{im} \cdot \frac{\partial}{\partial \mu_{im}} \left[\exp\left\{ -\frac{1}{2} (x_t - \mu_{im})^\intercal \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right\} \right]}{\sqrt{(2\pi)^K \cdot |\Sigma_{im}|} \cdot \sum_{m'=1}^{M} W_{im'} \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})} \cdot \langle Z_t \rangle_i \tag{25}$$

To compute the partial derivative in the numerator of equation (25), recall that

$$(x_t - \mu_{im})^\intercal \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) = tr\left\{ \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \cdot (x_t - \mu_{im})^\intercal \right\}$$

where $tr\{\cdot\}$ denotes the matrix trace which can be simplified to

$$tr\left\{ \Sigma_{im}^{-1} \cdot x_t \cdot x_t^\intercal \right\} + tr\left\{ \Sigma_{im}^{-1} \cdot \mu_{im} \cdot \mu_{im}^\intercal \right\} - 2 \cdot tr\left\{ \Sigma_{im}^{-1} \cdot \mu_{im} \cdot x_t^\intercal \right\}$$

since the trace is an additive operator. Therefore,

$$\begin{aligned}
& \frac{\partial}{\partial \mu_{im}} \left[(x_t - \mu_{im})^\intercal \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right] \\
= \; & \frac{\partial}{\partial \mu_{im}} \left[tr\left\{ \Sigma_{im}^{-1} \cdot x_t \cdot x_t^\intercal \right\} + tr\left\{ \Sigma_{im}^{-1} \cdot \mu_{im} \cdot \mu_{im}^\intercal \right\} - 2 \cdot tr\left\{ \Sigma_{im}^{-1} \cdot \mu_{im} \cdot x_t^\intercal \right\} \right] \\
= \; & 2\mu_{im} \cdot \Sigma_{im}^{-1} - 2x_t \cdot \Sigma_{im}^{-1} \\
= \; & -2 \cdot (x_t - \mu_{im}) \cdot \Sigma_{im}^{-1}.
\end{aligned}$$

Now the partial derivative in the numerator of equation (25) can be easily evaluated using the chain rule to arrive at

$$\begin{aligned}
& \frac{\partial}{\partial \mu_{im}} \left[\exp\left\{ -\frac{1}{2} (x_t - \mu_{im})^\intercal \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right\} \right] \\
= \; & \exp\left\{ -\frac{1}{2} (x_t - \mu_{im})^\intercal \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right\} \cdot (x_t - \mu_{im}) \cdot \Sigma_{im}^{-1}.
\end{aligned}$$

Therefore, equation (25) becomes

$$\sum_{t=1}^{T} \frac{W_{im} \cdot \mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \cdot (x_t - \mu_{im})}{\sum_{m'=1}^{M} W_{im'} \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})} \cdot \langle Z_t \rangle_i = 0$$

which can be written more conscisely in bracket notation as

$$\sum_{t=1}^{T} \langle Z_t X_t \rangle_{im} \cdot (x_t - \mu_{im}) = 0.$$

From here we can solve to obtain the otimal mean parameters,

$$\mu_{im}^* = \frac{\sum_{t=1}^{T} \langle Z_t X_t \rangle_{im} \cdot x_t}{\sum_{t=1}^{T} \langle Z_t X_t \rangle_{im}}$$

for $1 \leq i \leq N$ and $1 \leq m \leq M$.

Finally, it still remains to determine the optimal covariance parameters. Proceeding as above, computing the derivative of equation (24) with respect to the Σ_{im} and solving for 0, we get

$$0 = \sum_{t=1}^{T} \frac{W_{im} \cdot \frac{\partial}{\partial \Sigma_{im}} \left[\mathcal{N}(x_t; \mu_{im}, \Sigma_{im}) \right] \cdot \langle Z_t \rangle_i}{\sum_{m'=1}^{M} W_{im'} \cdot \mathcal{N}(x_t; \mu_{im'}, \Sigma_{im'})} \tag{26}$$

for all $1 \leq i \leq N$ and $1 \leq m \leq M$. We will consider

$$\frac{\partial \Sigma_{im}^{-1}}{\partial \Sigma_{im}}$$

as a matrix of partial derivatives with respect to the entries of Σ_{im}, where the jk^{th} entry is given by

$$\left(\frac{\partial \Sigma_{im}^{-1}}{\partial \Sigma_{im}} \right)_{jk} = \frac{\partial \Sigma_{im}^{-1}}{\partial \left(\Sigma_{im} \right)_{jk}}.$$

Since

$$\Sigma_{im} \cdot \Sigma_{im}^{-1} = \mathbb{I},$$

where $\mathbb{I}$ is the identity matrix, it follows from the product rule that

$$\frac{\partial \Sigma_{im}}{\partial \left(\Sigma_{im} \right)_{jk}} \cdot \Sigma_{im}^{-1} + \Sigma_{im} \cdot \frac{\partial \Sigma_{im}^{-1}}{\partial \left(\Sigma_{im} \right)_{jk}} = 0$$

and therefore

$$\frac{\partial \Sigma_{im}^{-1}}{\partial \left(\Sigma_{im} \right)_{jk}} = -\Sigma_{im}^{-1} \cdot e_j^{\intercal} \cdot e_k \cdot \Sigma_{im}^{-1} \tag{27}$$

where e_j and e_k are elementary row vector (i.e. e_j is equal to 0 in all but the j^{th} component, at which it is equal to 1). Using equation (27), we can compute the partial derivative

$$\begin{aligned}
& \frac{\partial}{\partial \left(\Sigma_{im} \right)_{jk}} \left[(x_t - \mu_{im})^{\intercal} \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right] \\
= \; & -(x_t - \mu_{im})^{\intercal} \cdot \Sigma_{im}^{-1} \cdot e_j^{\intercal} \cdot e_k \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \\
= \; & -\left(e_k \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \right) \cdot \left((x_t - \mu_{im})^{\intercal} \cdot \Sigma_{im}^{-1} \cdot e_j^{\intercal} \right) \\
= \; & -e_k \cdot \Sigma_{im}^{-1} \cdot (x_t - \mu_{im}) \cdot (x_t - \mu_{im})^{\intercal} \cdot \Sigma_{im}^{-1} \cdot e_j^{\intercal},
\end{aligned}$$

where the penultimate step works by splitting the expression into two scalars which necessarily commute. From here, we know that the full matrix of partial derivatives with respect to Σ_{im} is

$$\frac{\partial}{\partial \Sigma_{im}}\left[(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}\cdot(x_t-\mu_{im})\right] \;=\; -\Sigma_{im}^{-1}\cdot(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}.$$

Before we can fully compute out the partial derivative (26) we will need one additional fact from matrix algebra, namely

$$\frac{\partial\,|\Sigma_{im}|}{\partial\Sigma_{im}} = |\Sigma_{im}|\cdot\Sigma_{im}^{-1}$$

which is a consequence of Jacobi's formula, and from which we can deduce

$$\frac{\partial}{\partial\Sigma_{im}}\left[|\Sigma_{im}|^{-\frac{1}{2}}\right] = -\frac{1}{2}\,|\Sigma_{im}|^{-\frac{3}{2}}\cdot|\Sigma_{im}|\cdot\Sigma_{im}^{-1} = -\frac{1}{2}\,|\Sigma_{im}|^{-\frac{1}{2}}\cdot\Sigma_{im}^{-1}.$$

Now the partial derivative in the numerator of equation (26), can be easily computed using the product rule and the equations above to obtain

$$\begin{aligned}&\frac{\partial}{\partial\Sigma_{im}}\left[(2\pi)^{-\frac{K}{2}}\cdot|\Sigma_{im}|^{-\frac{1}{2}}\cdot\exp\left\{-\frac{1}{2}(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}\cdot(x_t-\mu_{im})\right\}\right]\\ =\;& -\frac{1}{2}\cdot\mathcal{N}(x_t;\mu_{im},\Sigma_{im})\cdot\Sigma_{im}^{-1}\cdot\left(1-(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}\right).\end{aligned}$$

Substituting this into equation (26) and dividing out the unnecessary terms, we get

$$\begin{aligned}0 \;&=\; \sum_{t=1}^{T}\frac{W_{im}\cdot\mathcal{N}(x_t;\mu_{im},\Sigma_{im})\cdot\left(1-(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}\right)\cdot\langle Z_t\rangle_i}{\sum_{m'=1}^{M}W_{im'}\cdot\mathcal{N}(x_t;\mu_{im'},\Sigma_{im'})}\\ &=\; \sum_{t=1}^{T}\langle Z_tX_t\rangle_{im}\cdot\left(1-(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal\cdot\Sigma_{im}^{-1}\right)\end{aligned}$$

which implies

$$\sum_{t=1}^{T}\langle Z_tX_t\rangle_{im}\cdot\Sigma_{im} = \sum_{t=1}^{T}\langle Z_tX_t\rangle_{im}\cdot(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal.$$

This can now be solved to obtain the optimal covariancec parameter

$$\Sigma_{im}^{*} \;=\; \frac{\sum_{t=1}^{T}\langle Z_tX_t\rangle_{im}\cdot(x_t-\mu_{im})\cdot(x_t-\mu_{im})^\intercal}{\sum_{t=1}^{T}\langle Z_tX_t\rangle_{im}} \tag{28}$$

for all $1\leq i\leq N$ and $1\leq m\leq M$.

3.4. **Expectation maximization for flexible hybrid state emissions.** In this section we will merge the discussions of the previous section to develop a maximally flexible framework for maximizing the expected value. In particular, we will allow for several disjoint observation sequences, also called *episodes*, and we will allow observations to be drawm from hybrid state distributions, consisting of both discrete and Gaussian probability distributions. Suppose that

$$\{X_{1:T_1},...,X_{1:T_E}\}$$

are E distinct observation episode, where episode e consists of T_e observations,

$$\{X_{e,1},...,X_{e,T_e}\}.$$

Suppose that for any episode, e, and time t, an observation vector is drawn from

$$X_{e,t} \sim \mathcal{X}_{\text{discrete}} \times \mathcal{X}_{\text{continuous}}$$

where X_{discrete} is a discrete space where a random variable can take on one of D distinct possible values, and $X_{\text{continuous}}$ is an M-component K-dimensional Gaussian mixture model.

Suppose the latent states corresponding to these observation episodes are given by

$$\{Z_{1:T_1}, ..., Z_{1:T_E}\}$$

where the hidden state at time t for episode e is $Z_{e,t}$. Since the episodes are disjoint in time we can maximize the expected value across episodes by additively maximizing the expected value in each individual episode. In this setting, an HMM is fully parameterized by

$$\theta = \{\pi, A, B, W, \mu, \Sigma\}$$

and the expected value is given by,

$$\begin{aligned}
\sum_{e=1}^{E} \mathbb{E}_{Z_{1:T_e}|X_{1:T_e},\theta'} \left[\log p(X_{1:T_e}, Z_{1:T_e} \mid \theta)\right] &= \sum_{e=1}^{E}\sum_{i=1}^{N} \log \pi_i \cdot \langle Z_{e,t}\rangle_i + ... \\
&... + \sum_{e=1}^{E}\sum_{t=2}^{T_e}\sum_{i=1}^{N}\sum_{j=1}^{N} \log A_{ij} \cdot \langle Z_{e,t-1}Z_{e,t}\rangle_{ij} + ... \\
&... + \sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{i=1}^{N}\sum_{d=1}^{D} \delta_{d,X_{e,t}} \log B_{di} \cdot \langle Z_{e,t}\rangle_i \\
&... + \sum_{e=1}^{E}\sum_{t=1}^{T_e}\sum_{i=1}^{N} \log \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_{e,t}; \mu_{im}, \Sigma_{im}) \cdot \langle Z_{e,t}\rangle_i,
\end{aligned}$$

which can be optimized using exactly the arguments laid out in the previous sections. In this way we arrive at the fully flexible update equations as follows,

$$\pi_i^* = \frac{1}{E}\sum_{e=1}^{E}\langle Z_{e,1}\rangle_i \tag{29}$$

$$A_{ij}^* = \frac{\sum_{e=1}^{E}\sum_{t=2}^{T_e}\langle Z_{e,t-1}Z_{e,t}\rangle_{ij}}{\sum_{e=1}^{E}\sum_{t=2}^{T_e}\sum_{j'=1}^{N}\langle Z_{e,t-1}Z_{e,t}\rangle_{ij'}} \tag{30}$$

$$B_{di}^* = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\delta_{d,X_{e,t}} \cdot \langle Z_{e,t}\rangle_i}{\sum_{e=1}^{E}\sum_{d'=1}^{D}\sum_{t=1}^{T_e}\delta_{d',X_{e,t}} \cdot \langle Z_{e,t}\rangle_i} \tag{31}$$

$$W_{im}^* = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}X_{e,t}\rangle_{im}}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}\rangle_i} \tag{32}$$

$$\mu_{im}^* = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}X_{e,t}\rangle_{im} \cdot x_{e,t}}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}X_{e,t}\rangle_{im}} \tag{33}$$

$$\Sigma_{im}^* = \frac{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}X_{e,t}\rangle_{im} \cdot (x_{e,t} - \mu_{im}) \cdot (x_{e,t} - \mu_{im})^\intercal}{\sum_{e=1}^{E}\sum_{t=1}^{T_e}\langle Z_{e,t}X_{e,t}\rangle_{im}} \tag{34}$$

for $1 \leq i, j \leq N$, $1 \leq d \leq D$ and $1 \leq m \leq M$.

Therefore, to recap, the goal of EM is to iteratively maximize the conditional expected value until we reach an optimal set of model parameters that maximize the likelihood. We can compute the expected value without too much difficult using the forward-backward algorithm (Algorithm 1), and from the discssions above, we can also maximize the expected value. Combinging all of this, we now give pseudocode for a flexible EM for hybrid state emissions in Algorithm 2.

Algorithm 2 Flexible EM Algorithm for Hybrid HMM

Input Model, θ; observations, $\{x_{1:T_1}, ..., x_{1:T_E}\}$; convergence threshold, ϵ
Output Model, θ^*, which maximizes likelihood for given observations .

function ExpectationMaximization($\theta, \{x_{1:T_1}, ..., x_{1:T_E}\}, \epsilon$)
  for all episodes $e \in \{1, ..., E\}$ **do**
    $\alpha_e, \beta_e \leftarrow$ ForwardBackward($\theta, X_{1:T_e}$)
    for all timesteps $t \in \{1, ..., T_e\}$ **do**
      $\langle Z_{e,t} \rangle$, $\langle Z_{e,t-1} Z_{e,t} \rangle$, $\langle Z_{e,t} X_{e,t} \rangle \leftarrow$ Compute using equations (14), (15), (22).
    end for
  end for
  $\theta^* \leftarrow$ Update model parameters using equations (29) - (34).
  if $\max\{|\theta - \theta^*|\} > \epsilon$ **then**
    $\theta \leftarrow \theta^*$
    Repeat steps 2 - 8
  end if
  return θ^*
end function

In practice, the probabilties in question will often be very close to 0, therefore using floating point numbers with limited precision, will often lead to undesireable over/underflow errors. Therefore it will be useful to confugure the algorithm in terms of the multivariate softplus function, or LogSumExp (LSE) function which will give a smooth approximation to the maximum value, defined as

$$\text{LSE}(y_1, ..., y_N) = y^* + \log(\exp(y_1 - y^*) + ... + \exp(y_N) - y^*))$$

where $y^* = \max\{y_1, ..., y_N\}$ (cf. [3]). Using this construction, we can easily compute logarithms of our bracket notation from equations (14), (15) and (22) and carry out slightly modified version of equations (29)-(34).

4. Flexible Inference and Imputation

Once equipped with an optimal set of model parameters for a series of observation episodes, we may wish to infer the most likely sequence of hidden states, for an episode, or impute missing observation values in other similarly structured data. These tasks of *inference* and *imputation* can be carried out using variants of Algorithm 1.

4.1. Inference of hidden states.

The most likely sequence of hidden states can be inferred using the Viterbi Algorithm, which we give below as Algorithm 3. This looks very similar to the forward algorithm, except as we pass through the forward step we store a record of the realtive probability of states visited, and use this to fill in the sequence of most likely hidden states. The $N \times T$-dimensional *Viterbi array*, which we denote V is filled analogously to the

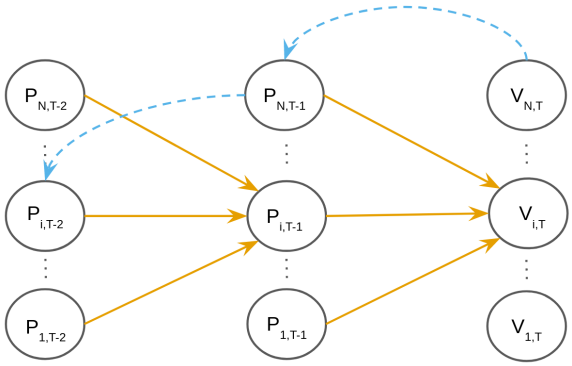


FIGURE 5. In the Viterbi algorithm, the sequence of most likely hidden states is traced beginning from time T and following the entries of the backpointer array.

forward algorithm, summing up the cumulative likelihood of arriving in hidden state h_i from all of the possible prior states (for state h_i, these are shown as solid orange lines in Figure 5). This gives

$$V_{i1} = \max_j \pi_i \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j)$$

and

$$V_{it} = \max_j V[j, t-1] \cdot A_{ji} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j).$$

Along the way, the most likely prior state is recorded in the $N \times T$-dimensional *backpointer array*, which we denote P, so

$$P_{i1} = 0$$

and

$$P_{it} = \operatorname*{argmax}_j V[j, t-1] \cdot A_{ji} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j).$$

After finishing the full foward pass, we begin at the end of our hidden state sequence, letting z_T be the most likely final state, that is,

$$z_T = \operatorname*{argmax}_i V_{iT}.$$

Continuing recursively from here, the t^{th} entry of the backpointer matrix indicates the most likely state that came prior whichever state was chosen as z_{t+1}, so we set z_t equal to the entry in row z_{t+1} and column t of P (This path is indicated as the dashed blue line in Figure 5). Pseudocode for this process is given in Algorithm 3.

4.2. **Imputation of missing observations.** Suppose we have a trained model and an additional sequence of observations, $x_{1:T}$, which was generated by a process analogous to the that which generated our training data. A common issue in dealing with real world data is the need to handle missing values. This can occur because of errors on signal processing, network downtime, or poor historical record keeping. The goal of imputation is to fill in missing values using statistically relevant data. In the case of an HMM, if we are missing

Algorithm 3 Flexible Viterbi Algorithm for Hybrid HMM

Input Model, θ; observations, $x_{1:T}$.
Output Most likely sequence of hidden states and its associated probability.

function VITERBI($\theta, x_{1:T}$)
  Initialize empty $N \times T$-dimensional Viterbi path array, V.
  Initialize empty $N \times T$-dimensional backpointer array, P.
  Initialize empty $T \times 1$ array of hidden states, Z.
  $V[i,1] \leftarrow \max_j \pi_i \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j)$
  $P[i,1] \leftarrow 0$
  for all timesteps $t \in \{1, ..., T\}$ **do**
    for all hidden states $i \in \{1, ..., N\}$ **do**
      $V[i,t] \leftarrow \max_j V[j,t-1] \cdot A_{ji} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j)$
      $P[i,t] \leftarrow \text{argmax}_j V[j,t-1] \cdot A_{ji} \cdot \prod_{k=1}^{K} p_k(x_t \mid h_j)$
    end for
  end for
  ViterbiProbability $\leftarrow \max_i V[i,T]$
  $Z[T] \leftarrow \text{argmax}_i V[i,T]$
  for all timesteps $t \in \{T-1, ..., 1\}$ **do**
    $Z[t] \leftarrow P[Z[t+1], t]$
  end for
  return Z, ViterbiProbability
end function

some values at a certain timestep, then we still have information about data observered prior to, subsequent to, and concurrent with the missing values. This information can be used to determine the best possible choice for the missing values.

Suppose $X_{1:T}$ is a sequence of observations, where x_t is only partially observed. We will let y denote the recorded observation at timestep t, and let the random variable Y^* denote the unknown portion of the observation, that is

$$X_t = y \oplus Y^*.$$

Our goal is to find the most likely Y^* given everything that we know, which we can define as

$$y^* = \underset{Y^*}{\text{argmax}}\, p(Y^* \mid X_{1:T}) = \underset{Y^*}{\text{argmax}} \sum_{i=1}^{N} p(Y^* \mid y, Z_t = h_i) \cdot p(Z_t = h_i \mid x_{1:t-1}, y, x_{t+1:T})$$

by marginalizing over all possible hidden states at time t. The second term on the right hand side above is the relative likelihood of each of the hidden states given everything that we know. We will write this more consisely using a slightly modified bracket notation, $\langle Z_t \rangle^*$, of which the i^{th} component denotes the probably of hidden state h_i given all of the observed data in $x_{1:T}$, that is

$$\langle Z_t \rangle_i^* \quad = \quad p(Z_t = h_i \mid x_{1:t-1}, y, x_{t+1:T}). \tag{35}$$

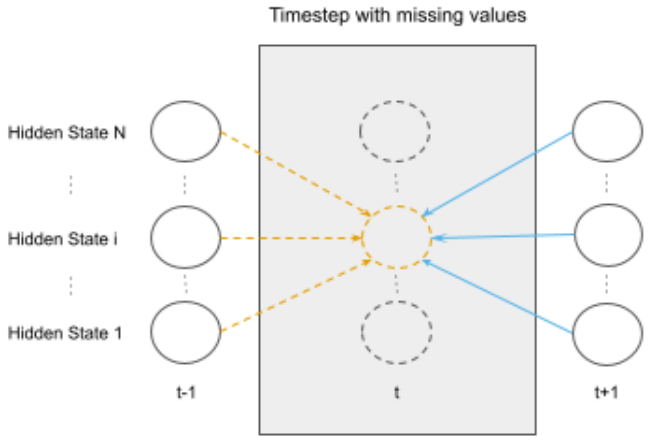


FIGURE 6. If data is only partially observed at time t, then we can directly compute the backward probabilty of any hidden state, indicated with the solid blue line, but we can only compute forward probabilities up to timestep $t-1$ combined with transition and emission probabilties, indicated with the dashed orange line.

Once again the Markov property allows us to simplify this expression without too much difficulty, since

$$p(Z_t = i, x_{1:t-1}, y, x_{t+1:T}) \quad = \quad p(y \mid Z_t = h_i) \cdot \beta_t(h_i) \cdot \sum_{j=1}^{N} \alpha_{t-1}(h_j) \cdot A_{ji}.$$

As shown in Figure 6, we can directly apply our knowledge of everything subsequent to time t by computing the backward probability (the solid blue line). From the forward direction, we can only compute the probability up until timestep $t-1$, and then compute transition and emission probabilities by hidden state, indicated by the dashed orange line.

Thus, we need to compute

$$\langle Z_t \rangle_i^* = \frac{p(y \mid Z_t = h_i) \cdot \beta_t(h_i) \cdot \sum_{j=1}^{N} \alpha_{t-1}(h_j) \cdot A_{ji}}{\sum_{i'=1}^{N} p(y \mid Z_t = h_{i'}) \cdot \beta_t(h_{i'}) \cdot \sum_{j=1}^{N} \alpha_{t-1}(h_j) \cdot A_{ji'}}, \tag{36}$$

all of which we have already seen, except perhaps, $p(y \mid Z_t = h_i)$. The undrlying probability distribution of the components of the observation will be relevent, therefore we will decompose Y^* as

$$Y^* = (Y^*_{dis}, Y^*_{gmm})$$

where Y^*_{dis} and Y^*_{gmm} are drawn from discrete and Gaussian mixture model distributions, repsectively, and we will decompose y similarly. In this way, the full vector at time t is given by

$$x_t = (y_{dis}, y_{gmm}, Y^*_{dis}, Y^*_{gmm}).$$

Therefore, our quantity of interest, becomes

$$p(y \mid Z_t = h_i) \quad = \quad p(y_{dis} \mid Z_t = h_i) \cdot p(y_{gmm} \mid Z_t = h_i). \tag{37}$$

For the discrete part, recall that (y_{dis}, Y^*_{dis}) can take on one of D distinct discrete values. The conditional probabilty of y_{dis} can be obtained by summing over any of those values which match y_{dis} in the relevant components. By a slightly abuse of the δ notation (taken to mean, the quantities are equal in all of the known components), we can write this as

$$p(y_{dis} \mid Z_t = h_i) \;=\; \sum_{d=1}^{D} \delta_{d,y_{dis}} \cdot b_{di}. \tag{38}$$

For the continuous part, we know that since (y_{gmm}, Y^*_{gmm}) is drawn from a Gaussian mixure model, y_{gmm} is also drawn from a Gaussian mixture model. In the interest of thoroughness we will prove that below in Lemma 4.1, and since it will be useful, we will also give an explicit form for the conditional probability of Y^*_{gmm} given y_{gmm}.

Lemma 4.1. *Let χ be a K-dimensional Gaussian mixture model with M components having weights W_m, means μ_m and covariance matrices Σ_m for $1 \leq m \leq M$. For a vector $(x_a, x_b) \sim \mathcal{X}$ we can decompose the means and covariances as*

$$\mu_m = \begin{bmatrix} \mu_{m,a} \\ \mu_{m,b} \end{bmatrix} \qquad \textit{and} \qquad \Sigma_m = \begin{bmatrix} \Sigma_{m,aa} & \Sigma_{m,ab} \\ \Sigma_{m,ba} & \Sigma_{m,bb} \end{bmatrix}.$$

Then the conditional probability of x_b given x_a is

$$p(x_b \mid x_a) = \sum_{m=1}^{M} W_m \frac{\mathcal{N}(x_a, x_b; \mu_m, \Sigma_m)}{\mathcal{N}(x_a; \mu_{m,a}, \Sigma_{m,a})} = \sum_{m=1}^{M} W_m \cdot \mathcal{N}(x_b; \mu^*_m, \Sigma^*_m),$$

and the marginal probability of x_a is

$$p(x_a) = \sum_{m=1}^{M} W_m \cdot \mathcal{N}(x_a; \mu_{m,aa}, \Sigma_{m,aa}) \tag{39}$$

where

$$\mu^*_m = \mu_{m,b} - \Sigma_{m,ba} \cdot \Sigma^{-1}_{m,aa} \left(x_a - \mu_{m,a}\right) \tag{40}$$

and

$$\Sigma^*_m = \Sigma_{m,bb} - \Sigma_{m,ba} \cdot \Sigma^{-1}_{m,aa} \cdot \Sigma_{m,ab}. \tag{41}$$

Proof. Let $\mathcal{X}$ be a K-dimensional Gaussian mixture model. Since the probability density function is linear across the Gaussian mixture model components we will omit the m subscript from this proof in the interest of readability. First we will recall that the integral over a probability density function is equal to 1, so

$$\int \mathcal{N}(x; \mu, \Sigma)\; d\mathcal{X} = 1, \tag{42}$$

from which we know

$$\sqrt{(2\pi)^K \cdot |\Sigma|} = \int \exp\left\{-\frac{1}{2}(x-\mu)^\intercal \cdot \Sigma^{-1} \cdot (x-\mu)\right\}\; d\mathcal{X}.$$

For $x \sim \mathcal{X}$, decompose x as (x_a, x_b), where $a+b=K$, decompose the means and covariances as

$$\mu = \begin{bmatrix} \mu_a \\ \mu_b \end{bmatrix} \qquad \text{and} \qquad \Sigma = \begin{bmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{bmatrix},$$

and define

$$\Sigma^{-1} = V = \begin{bmatrix} V_{aa} & V_{ab} \\ V_{ba} & V_{bb} \end{bmatrix}.$$

We will begin by pointing out some elementary identities in involving block matrices. First, we know that Σ can be inverted blockwise in such a way that

$$V_{bb} = \left(\Sigma_{bb} - \Sigma_{ba} \cdot \Sigma_{aa}^{-1} \cdot \Sigma_{ab}\right)^{-1} \tag{43}$$

and

$$\Sigma_{aa} = \left(V_{aa} - V_{ab} \cdot V_{bb}^{-1} \cdot V_{ba}\right)^{-1}, \tag{44}$$

moreover, since Σ_{aa} is invertible, we also know that

$$|\ \Sigma\ | = |\ \Sigma_{aa}\ | \cdot |\ \Sigma_{bb} - \Sigma_{ba} \cdot \Sigma_{aa}^{-1} \cdot \Sigma_{ab}\ |. \tag{45}$$

The probably of x_a can be marginalized as

$$\mathcal{N}(x_a; \mu_{aa}, \Sigma_{aa}) = \int \mathcal{N}(x_a, x_b; \mu, \Sigma)\ dx_b.$$

or equivalently

$$\mathcal{N}(x_a; \mu_{aa}, \Sigma_{aa}) = \frac{1}{\sqrt{(2\pi)^K \cdot |\Sigma|}} \int \exp\left\{-\frac{1}{2} \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}^\intercal \cdot \Sigma^{-1} \cdot \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}\right\}\ dx_b.$$

Using a higher dimensional analogue of completing the square, we can simply the matrix product inside the exponential in the previous equation, to

$$\begin{aligned} & \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}^\intercal \cdot \Sigma^{-1} \cdot \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix} \\ = \ & (x_a - \mu_a)^\intercal \cdot \left(V_{aa} - V_{ab} \cdot V_{bb}^{-1} \cdot V_{ba}\right) \cdot (x_a - \mu_a) + ... \\ & ... + \left(V_{bb}^{-1} \cdot V_{ba} \cdot (x_a - \mu_a) + (x_b - \mu_b)\right)^\intercal \cdot V_{bb} \cdot \left(V_{bb}^{-1} \cdot V_{ba} \cdot (x_a - \mu_a) + (x_b - \mu_b)\right). \end{aligned}$$

Since Σ is a block matrix, we have some helpful identities involving the inversion of its components, in particular, it is known that

$$\Sigma_{aa}^{-1} = V_{aa} - V_{ab} \cdot V_{bb}^{-1} \cdot V_{ba}.$$

We will define as

$$\hat{\mu} = \mu_b - V_{bb}^{-1} \cdot V_{ba} \cdot (x_a - \mu_a). \tag{46}$$

Then the marginalized probability can be written as

$$\mathcal{N}(x_a; \mu_{aa}, \Sigma_{aa}) = \frac{\exp\left\{-\frac{1}{2}(x_a - \mu_a)^\intercal \cdot \Sigma_{aa}^{-1} \cdot (x_1 - \mu_a)\right\}}{\sqrt{(2\pi)^K \cdot |\Sigma|}} \int \exp\left\{-\frac{1}{2}(x_b - \hat{\mu})^\intercal \cdot V_{bb} \cdot (x_b - \hat{\mu})\right\}\ dx_b,$$

which can be further simplified to

$$\mathcal{N}(x_a; \mu_{aa}, \Sigma_{aa}) = \frac{\exp\left\{-\frac{1}{2}(x_a - \mu_a)^\intercal \cdot \Sigma_{aa}^{-1} \cdot (x_1 - \mu_a)\right\}}{\sqrt{(2\pi)^a \cdot |\Sigma_{aa}|}}, \tag{47}$$

in light of the equations (42), (44) and (45), thereby confirming equation 39. Now we can compute the conditional probabilty,

$$p(x_b \mid x_a) = \frac{\mathcal{N}(x_a, x_b; \mu, \Sigma)}{p(x_a)} = \frac{\exp\left\{-\frac{1}{2}\begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}^{\intercal} \cdot \Sigma^{-1} \cdot \begin{bmatrix} x_a - \mu_a \\ x_b - \mu_b \end{bmatrix}\right\}}{\sqrt{(2\pi)^b \cdot \left|V_{bb}^{-1}\right|} \cdot \exp\left\{-\frac{1}{2}(x_a - \mu_a)^{\intercal} \cdot \Sigma_{aa}^{-1} \cdot (x_1 - \mu_a)\right\}}.$$

Using the same calculations as above this can be simplified to

$$p(x_b \mid x_a) = \frac{\mathcal{N}(x_a, x_b; \mu, \Sigma)}{p(x_a)} = \frac{\exp\left\{-\frac{1}{2}(x_b - \mu^*)^{\intercal} \cdot V_{bb}^{-1} \cdot (x_b - \mu^*)\right\}}{\sqrt{(2\pi)^b \cdot \left|V_{bb}^{-1}\right|}},$$

where μ^* is exactly as in equation (40), as desired. Furthermore,

$$\Sigma^* = V_{bb}^{-1} = \Sigma_{bb} - \Sigma_{ba} \cdot \Sigma_{aa}^{-1} \cdot \Sigma_{ab},$$

which is equation (41). Moreover, combining this with equation (47) we get equation 39. □

With this, we have

$$p(y_{gmm} \mid Z_t = h_i) \quad = \quad \sum_{m=1}^{M} W_{im} \cdot \mathcal{N}(x_a; \mu_{im,gmm}, \Sigma_{im,gmm}) \tag{48}$$

where $N(x_a; \mu_{im,gmm}, \Sigma_{im,gmm})$ is as in equation (39) from Lemma 4.1. Therefore, combining equations (38) and (48) with equation (37), we have all of the necessary ingredients to compute the vector of relative probabilities for the hidden states $\langle Z_t \rangle^*$.

Now we are ready to impute. For the discrete portion, using equation (38) and an application of Bayes' theorem, we obtain

$$p(Y_{dis}^* \mid y, Z_t = i) = \frac{b_{(y_{dis}, Y_{dis}^*)i}}{\sum_{d=1}^{D} \delta_{d,y_{dis}} \cdot b_{di}}.$$

Thus the most likely value for our missing discrete data is

$$y_{dis}^* = \underset{d}{\operatorname{argmax}} \sum_{i=1}^{N} \frac{b_{(y_{dis}, Y_{dis}^*)i}}{\sum_{d=1}^{D} \delta_{d,y_{dis}} \cdot b_{di}} \cdot \langle Z_t \rangle_i^*. \tag{49}$$

We now have several options for imputing Y_{gmm}^*. One method, which we'll call `argmax` is to impute with the mean of the most likely hidden state and mixture component. By setting

$$h = \underset{i}{\operatorname{argmax}} \langle Z_t \rangle_i^*$$

and the most likley mixture component

$$c = \underset{m}{\operatorname{argmax}} W_{hm}$$

and let y_{gmm}^* be equal to the mean value of the most likley mixture component from the most likely hidden state,

$$y_{gmm}^* \quad = \quad \mu_{hc}^*.$$

where μ_{hc}^* is defined as in equation (40). Another option, which we'll call `maximal` is to impute with the most probable mean. That is, let

$$\mu_{im}^* \quad = \quad \underset{i,m}{\operatorname{argmax}} \langle Z_t \rangle_i^* \cdot W_{im} \cdot \mathcal{N}(\mu_{im}; \mu_{im}, \Sigma_{im})$$

and then

$$y^*_{gmm} \quad = \quad \mu^*_{im}.$$

Finally, for the `average` method, we will take the full weighted sum across all hidden states

$$y^*_{gmm} \quad = \quad \sum_{i=1}^{N} \sum_{m=1}^{M} W_{im} \mu^*_{im} \cdot \langle Z_t \rangle^*_i.$$

Depending on the underlying data and the goal of the imputation, these methods can have varying utility.

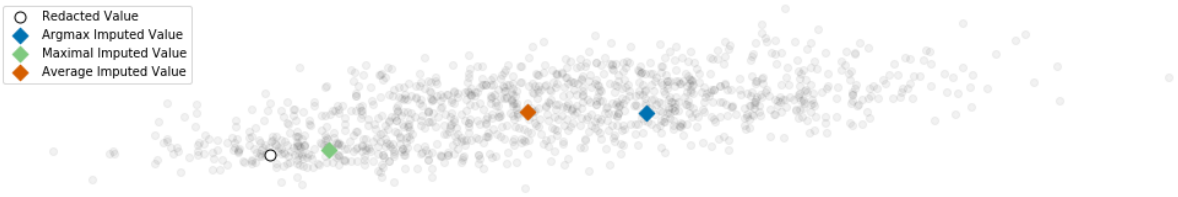


FIGURE 7. In this example, the circled data point was artificially redacted in order to demonstrate the utility of the three imputation methods, `argmax`, `maximal`, and `average`

In Figure 7, we've artificially redacted a value in the synthetic data presented in Figure 1. Using the three methods of imputation described above, we are able to acheive a best guess for this missing value. Notice that the dark blue value, which is obtained by the first method is relatively far from the missing data point. This is because the `argmax` will always be biased towards hidden states that occur in large quantity. This means that it's difficult to impute rare values, but on the other hand the imputed value is guaranteed to be statistically meaningful. The `maximal` method on the other hand, shown here in green, is quite successful at imputing this missing value, even though it's coming from a cluster that has a proportionally smaller share of values. This is a consequence of the fact that this method favors densely clustered points with high probability. Finally, we see that the `average` method also performs poorly at the task of imputing the correct missing value. This method will rarely impute correctly, but is limited in how far it can drift from the correct value. The general strategy for imputation is to use `maximal` when the goal is to be the most correct, but to use `average` or `argmax` when the goal is to be less wrong.

REFERENCES

[1] C. M. Bishop, *Pattern recognition and machine learning*, Springer, 2006.
[2] Z. Ghahramani and M. I. Jordan, Factorial hidden markov models, *Machine Learning*, **29** (1997), pp. 245–273.
[3] F. Nielsen and K. Sun, Guaranteed bounds on the kullback–leibler divergence of univariate mixtures, *IEEE Signal Processing Letters*, **23** (2016), pp. 1543–1546.

DATA INTENSIVE STUDIES CENTER, TUFTS UNIVERSITY, MEDFORD, MA 02155
Email address: `anna.haensch@tufts.edu`