Homework 1 Responses

Chris Crabtree

Sept. 30, 2021

I spent a VERY long time figuring out #2. I did not have time to try #1 or the bonus.

1

2

We use the following stochastic process as our data generation model for the time dependent variable y_t in an arbitrary time step $t \in [1, T]$. Note that, except for the y_t's, all of the following are random variables generated by parametric distributions defined by our model. The stochastic process is modelled as follows:

$$
\begin{aligned}
\mathbf{z}_1 &= \mathbf{s}_0 \\
y_1 &= \mathbf{x}_1{}^\intercal \mathbf{z}_1 \\
\mathbf{z}_2 &= \mathbf{Z}_2^* \cdot \mathbf{z}_1 \\
y_2 &= \mathbf{x}_2{}^\intercal \mathbf{z}_2 \\
\mathbf{z}_3 &= \mathbf{Z}_3^* \cdot \mathbf{z}_2 \\
&\vdots \\
\mathbf{z}_{t-1} &= \mathbf{Z}_{t-1}^* \cdot \mathbf{z}_{t-2} \\
y_{t-1} &= \mathbf{x}_{t-1}{}^\intercal \mathbf{z}_{t-1} \qquad (1) \\
\mathbf{z}_t &= \mathbf{Z}_t^* \cdot \mathbf{z}_{t-1} \qquad (2) \\
y_t &= \mathbf{x}_t{}^\intercal \mathbf{z}_t \qquad (3)
\end{aligned}
$$

With $\mathbf{z}_t$ and the columns of $\mathbf{Z}_t^*$, $t \in [1, T]$, being random standard basis vectors in $\mathbb{R}^K$. As such, I will use z_t to denote the index of the non-zero entry in $\mathbf{z}_t$ and likewise $\mathbf{z}_{tj}^*{}^\intercal$ to denote the j^{th} column of $\mathbf{Z}_t^*$.

I will also use the notation $\mathbf{z}_{1:t} \in \mathbb{R}^{K \times t}$ to refer to the sequence of $\mathbf{z}_t$ vectors chosen on the left side of the equalities in the above generation. $\mathbf{Z}_t^* \in \mathbb{R}^{K \times K}$ is akin to a set of candidate $\mathbf{z}_t$'s. From the above sequence, it should be clear that $\mathbf{z}_t$ determines both $\mathbf{y}_t$ and $\mathbf{z}_{t+1}$. $\mathbf{z}_t$ is effectively the ${z_{t-1}}^{th}$

column of $\mathbf{Z}_t^*$. $\mathbf{a}_0$ and $\mathbf{Z}_t^*$ are distributed as follows:

$$\mathbf{s}_0 \sim \text{Multi}(1, \boldsymbol{\pi}_0), \boldsymbol{\pi}_0 \in \mathbb{R}^K \tag{4}$$

$$\boldsymbol{\pi}_0 \sim \text{Dir}(1/K, 1/K, \cdots, 1/K) \tag{5}$$

$$\mathbf{Z}_t^* = \begin{bmatrix}\mathbf{z}_{t1}^* & \mathbf{z}_{t2}^* & \cdots & \mathbf{z}_{tK}^*\end{bmatrix} \sim \begin{bmatrix}\text{Multi}(1, \boldsymbol{\pi}_1) & \text{Multi}(1, \boldsymbol{\pi}_2) & \vdots & \text{Multi}(1, \boldsymbol{\pi}_K)\end{bmatrix}, \ \boldsymbol{\pi}_i \in \mathbb{R}^K \tag{6}$$

$$\mathbf{P} = \begin{bmatrix}\pi_{11} & \pi_{12} & \cdots & \pi_{1K} \\ \pi_{21} & \pi_{22} & \cdots & \pi_{2K} \\ \vdots & & \cdots & \vdots \\ \pi_{K1} & \pi_{K2} & \cdots & \pi_{KK}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\pi}_1{}^\intercal \\ \boldsymbol{\pi}_2{}^\intercal \\ \vdots \\ \boldsymbol{\pi}_K{}^\intercal\end{bmatrix} \overset{iid}{\sim} \text{Dir}(1/K, 1/K, \cdots, 1/K) \tag{7}$$

$$\tag{8}$$

$\mathbf{x}_t$ is defined as follows with $\mathbf{P}$ having the same definition as above:

$$\mathbf{x}_t = \begin{bmatrix}x_{t1} \\ x_{t2} \\ \vdots \\ x_{tK}\end{bmatrix} \overset{iid}{\sim} MVN(\boldsymbol{\mu}, \boldsymbol{\sigma}^{2\intercal}\mathbf{I}) = \begin{bmatrix}\mathcal{N}(\mu_1, \sigma_1^2) \\ \mathcal{N}(\mu_2, \sigma_2^2) \\ \vdots \\ \mathcal{N}(\mu_K, \sigma_K^2)\end{bmatrix} \tag{9}$$

$$\mu_j \overset{iid}{\sim} \text{Normal-Inv-Gamma}(\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2) \tag{10}$$

$$\tag{11}$$

We are interested in the posterior. I will denote it as:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P} | y_{1:t}, \mu_0, \sigma_0, \alpha_0, \beta_0) \tag{12}$$

Bayes theorem then gives us:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P} | y_{1:t}, \mu_0, \sigma_0, \alpha_0, \beta_0) = \frac{p(y_{1:t} | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0) p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P} | \mu_0, \sigma_0, \alpha_0, \beta_0)}{\int_{\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}} p(y_{1:t} | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0) p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P} | \mu_0, \sigma_0, \alpha_0, \beta_0)} \tag{13}$$

$$\propto p(y_{1:t} | \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0) p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P} | \mu_0, \sigma_0, \alpha_0, \beta_0) \tag{14}$$

First I will concentrate on the likelihood. Since we use the HMM model for time dependency, the likelihood can be decomposed simply. I will omit the priors to simplify notation (they are there, but won't be involved in any likelihood calculations). The likelihood is now:

$$
\begin{aligned}
p(y_{1:T}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) &= \sum_{\mathbf{z}_{1:T}} p(y_{1:T}|\mathbf{z}_{1:T}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) p(\mathbf{z}_{1:T}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) \\
&= \sum_{j=1}^{K} p(y_T|z_T = j, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) p(z_T = j | z_{(T-1)}, \mathbf{P}) \times \\
&\qquad \sum_{\mathbf{z}_{1:(T-1)}} p(y_{1:(T-1)}|z_{1:(T-1)}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \qquad \text{(HMM assumption)} \\
&= \sum_{j=1}^{K} p(y_T|z_T = j, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \times \\
&\qquad \sum_{i=1}^{K} p(z_T = j | z_{(T-1)} = i, \mathbf{P}) \sum_{\mathbf{z}_{1:(T-1)}} p(y_{1:(T-1)}|z_{1:(T-1)}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) \\
&= \sum_{j=1}^{K} p(\mathbf{x}^\intercal \mathbf{z}_T | z_T = j, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \times \\
&\qquad \sum_{i=1}^{K} p(\mathbf{z}_T = \mathbf{Z}_T^* \cdot \mathbf{z}_{T-1} \Rightarrow z_T = j | z_{(T-1)} = i, \mathbf{P}) \sum_{\mathbf{z}_{1:(T-1)}} p(y_{1:(T-1)}|z_{1:(T-1)}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) \qquad (15) \\
&= \sum_{j=1}^{K} \mathcal{N}(y_t|\mu_i, \sigma_i^2) \times \sum_{i=1}^{K} \boldsymbol{\pi}_{ij} \sum_{\mathbf{z}_{1:(T-1)}} p(y_{1:(T-1)}|z_{1:(T-1)}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) \qquad (16) \\
& \qquad (17)
\end{aligned}
$$

This summation is effectively over every possible combination of z_ts through time. A recursive definition is used here to allow for easier computation. Since the last factor is independent of both the j indexer we only need to store the K values for the $T-1$ time step. Note that the notation $\mathcal{N}(y_t|\mu_{z_t}, \sigma_{z_t}^2)$ refers to the density of y_t using the Normal distribution given those parameters. This is for clarification purposes. We can use the density here because the posterior we are calculating is acually a density despite the $p(\cdots)$ notation.

$$
\begin{aligned}
& p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}|\mu_0, \sigma_0, \alpha_0, \beta_0) \qquad (18) \\
& \qquad \propto \prod_{j=1}^{K} \text{Normal-Inv-Gamma}((\mu_j, \sigma_j^2)|\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2) \times \prod_{i=1}^{K} \text{Dir}(\boldsymbol{\pi}_i|1/K, 1/K, \cdots, 1/K) \qquad (19) \\
& \qquad (20)
\end{aligned}
$$

I separated the product for clarity. Each element of the K elements of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$ were generated once and likewise each $\boldsymbol{\pi_i} \in \mathbf{P}$ was generated once. With this we can finally write the posterior as:

$$p(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}|y_{1:t}, \mu_0, \sigma_0, \alpha_0, \beta_0) \tag{21}$$

$$\propto \tag{22}$$

$$p(y_{1:t}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0) \times p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}|\mu_0, \sigma_0, \alpha_0, \beta_0) \tag{23}$$

$$= \tag{24}$$

$$p(y_{1:t}|z_{1:t}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0)p(z_{1:t}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \mu_0, \sigma_0, \alpha_0, \beta_0) \times p_0(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}|\mu_0, \sigma_0, \alpha_0, \beta_0) \tag{25}$$

$$\propto \tag{26}$$

$$\sum_{j=1}^{K} \mathcal{N}(y_t|\mu_i, \sigma_i^2) \times \sum_{i=1}^{K} \boldsymbol{\pi}_{ij} \sum_{\mathbf{z}_{1:(T-1)}} p(y_{1:(T-1)}|z_{1:(T-1)}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{P}, \boldsymbol{\pi_0}) \tag{27}$$

$$\times \tag{28}$$

$$\prod_{j=1}^{K} \text{Normal-Inv-Gamma}((\mu_j, \sigma_j^2)|\mu_0, \sigma_0^2/\kappa_0, \nu_0, \sigma_0^2) \times \prod_{i=1}^{K} \text{Dir}(\boldsymbol{\pi}_i|1/K, 1/K, \cdots, 1/K) \tag{29}$$

$$\tag{30}$$

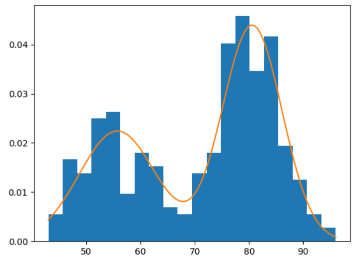
Now we need the conditionals of $\boldsymbol{\mu}$, $\boldsymbol{\sigma^2}$, and $\boldsymbol{P}$ to be able to use Gibbs sampling. We must include the $\mathbf{z}_{1:t}$'s in our conditioning to make the computation tractable. We start with the $\mathbf{z}_{1:t}$'s. We can ignore the prior since the $\mathbf{z}_{1:t}$'s do not appear there. Looking at Eq. 15 it should be clear that each $\mathbf{z}_t$ can be sampled iteratively using the recursive definition provided. Each $\mathbf{z}_t$ is involved in emitting y_t and in choosing $\mathbf{z}_{t+1}$. I used the Forward algorithm to do this. One note is that I computed everything in log-space for numerical stability rather than the telescoping normalization.

Now for $\boldsymbol{\pi}_j^\intercal \in \mathbf{P}$'s. For this we must think about what affects each $\boldsymbol{\pi}_j^\intercal$ in the posterior. The prior must be included, clearly, but the likelihood needs special attention. In particular, in Eq. 15 we can notice that a $\pi_{:j}$ only appears in Eq. 16 when it is chosen by $\mathbf{z}_{t+1}$. Since the $\mathbf{z}_{1:T}$s are given, we can simply count the number of $z_{t-1} = i$ and $z_t = j$ for all $t \in [1, T]$ and $i \in [1, K]$. Each time that occurs a π_{ij} will appear in the likelihood. Therefore we have:

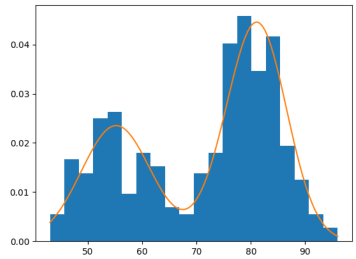
$$p(\boldsymbol{\pi}_j|-) \propto \prod_{t=1}^{T} \pi_{ij}^{\mathbb{1}(z_{t-1}=i \wedge z_t=j)} \text{Dir}(\boldsymbol{\pi}_j|1/K, 1/K, \cdots, 1/K) \tag{31}$$

The μ and σ^2 parameters are drawn from a Normal-Inv-Gamma distribution since it is conjugate to our likelihood with unknown μs and σ^2s. I used the parameter updates described in module 5 slide 6. The only difference is that for each component j, I only incorporate the datapoints from time steps where $z_t = j$.

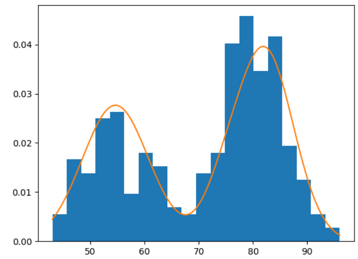
Figure 1 shows the results of fitting K component mixtures. I did not have time to complete the other plots, but they should be easy compared to the actual algorithm. For the stationary distribution I computed the eigendecomposition of the $\mathbf{P}$ matrix and found the eigenvector corresponding the eigenvalue of 1. Then I just normalized that eigenvector.



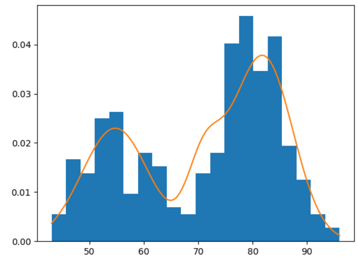
(a) K=2



(b) K=3



(c) K=4



(d) K=5

Figure 1: Estimated density using Gibbs sampled μs and σ^2's and the stationary distribution from the largest eigenvector. I did not have time to figure out how python can plot the credible intervals, although I have all samples so it shouldn't be hard.