A Variational Approach for Bayesian Density Regression

Eric Chuu ERICCHUU@STAT.TAMU.EDU
Department of Statistics
Texas A&M University
College Station, TX 77840, USA

Abstract

In the Bayesian density regression problem, a mixture of experts model is often used because of the high flexibility in estimating conditional densities. In this paper, we discuss the case when covariate dependent weights are used in the approximating mixture density. Under this framework, however, traditional Bayesian methods result in computational difficulties when the dimension of the covariates is large. In order to remedy this problem and to provide a method for faster inference, we propose using a variational approximation to estimate the conditional density. We also discuss upper bounds for approximating quantities that lack a closed form so that a coordinate ascent algorithm is viable.
Keywords: Bayesian Density Regression, Variational Bayes, Mixture Models

1. Introduction

In the Bayesian density regression problem, we observe data $(y_n, x_n)_{n=1}^N$, and the goal is the estimate the conditional density of $y \,|\, x$. A common approach for doing this is to model the density using a mixture of gaussians, such as the following,

$$f(y \mid x) = \sum_k \pi_k \mathcal{N}\left(y \mid \mu_k(x), \tau_k^{-1}\right) \tag{1}$$

While the representation of the density using predictor-independent weights yields less expensive computation, this approach often lacks flexibility to make it useful in practice and results in a reliance on have too many mixture components. As a result, there have been many proposed models that consider predictor-dependent weights. Some examples include using a kernel stick-breaking process (**dunsonpark:08**) or logit stick-breaking prior (**durante:17**) to generate the covariate-dependent weights. In the former method, the increased flexibility comes at heavy computational cost, and in the latter method, the process from which the weights are generated does not allow for intuitive inference. In our proposed model, the covariates enter the weights through a logistic link function so that we can naturally perform inference on the coefficients. More specifically, we can model

$$f(y \mid x) = \sum_k^K \pi_k(x) \mathcal{N}\left(y \mid \mu_k(x), \tau_k^{-1}\right) \tag{2}$$

where and $\pi_k \propto \exp(x^\intercal \gamma_k)$. In order to perform fast inference on the model parameters, we adopt a variational approach to obtain an approximating distribution to the true posterior.

Using this covariate-dependent setup, however, introduces a problem in the traditional coordinate ascent algorithm such that it prevents closed form updates of the variational distributions. In section 2, we give an overview of the priors used in the problem. In section 3, we consider the family of variational distributions that we use to approximate the true posterior. We then propose a way to obtain closed form updates by considering an upper bound on the problematic quantity, and in section 4, we formulate the complete algorithm. Finally, we discuss potential shortcomings of the proposed algorithm and other bounds that could be used to obtain more accurate approximating distributions.

2. Notation and Prior Specification

For the set of observed data, we denote $\mathbf{y} = \{y_1, \ldots, y_N\}, \mathbf{X} = \{x_1, \ldots, x_N\}$, where each $x_n \in \mathbb{R}^D$. Let $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$ and $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_k\}$ denote the D-dimensional coefficient vectors in the mixture weights and in gaussian mixture components, respectively. Finally, let $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_k\}$ denote the precision parameters. The mixture density can then be written explicitly as

$$p\left(y_n \mid x_n, \boldsymbol{\beta}, \boldsymbol{\tau}\right) = \sum_k \pi_k(x_n) \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right), \quad \pi_k(x_n) = \frac{\exp\{x_n^\intercal \gamma_k\}}{\sum\limits_{j=1}^{K} \exp\{x_n^\intercal \gamma_j\}} \tag{3}$$

We can simplify the form of the density by introducing the set of latent variables $\mathbf{Z} = \{z_1, \ldots, z_N\}$, where $z_n \in \mathbb{R}^K$ and $z_{nk} = 1$ if and only if y_n belongs to the k-th cluster so that $\sum_k z_{nk} = 1$. Conditioning on this additional variable $\mathbf{Z}$, we have the following density,

$$p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) = \prod_n \prod_k \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right)^{z_{nk}} \tag{4}$$

$$p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) = \prod_n \prod_k \pi_k(x_n)^{z_{nk}} = \prod_n \prod_k \left(\frac{\exp\{x_n^\intercal \gamma_k\}}{\sum_{j=1}^{K} \exp\{x_n^\intercal \gamma_j\}} \right)^{z_{nk}} \tag{5}$$

Next, we introduce the priors over the parameters $\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}$, where we simplify the calculations by considering conjugate priors. For ease of computation, we consider an independent standard normal prior on γ_k's, given by

$$p(\boldsymbol{\gamma}) = \prod_k p(\gamma_k) = \prod_k \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right) \tag{6}$$

For $(\boldsymbol{\beta}, \boldsymbol{\tau})$, we consider an independent normal-gamma prior, given by

$$p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right) = p\left(\boldsymbol{\beta} \mid \boldsymbol{\tau}\right) = \prod_k \prod_k \mathcal{N}\left(\beta_k \mid m_0, (\tau_k \Lambda_0)^{-1}\right) \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right) \tag{7}$$

Note that in the case where the mixture weights are covariate-independent, a Dirichlet prior is typically used for the mixing weights, $\pi_1, \ldots, \pi_K$. In this case, however, the mixing weights are fully specified by $\mathbf{X}$ and $\boldsymbol{\gamma}$, so we need only place a prior on $\boldsymbol{\gamma}$.

3. Variational Distribution

At this point, the variational parameters of interest are $\boldsymbol{\theta} = (\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma})$. The log of the joint distribution of these random variables (written up to constants) is given by

$$
\begin{aligned}
\ln p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) &= \ln\left\{p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \boldsymbol{\gamma}\right) p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)\right\} \\
&= \sum_n \sum_k z_{nk}\left\{-\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\tau_k - \frac{\tau_k}{2}\left(y_n - x_n^\intercal \beta_k\right)^2\right\} \\
&+ \sum_n \sum_k z_{nk}\left\{x_n^\intercal \gamma_k - \ln\sum_{j=1}^{K}\exp\{x_n^\intercal \gamma_j\}\right\} + \sum_k\left\{-\frac{D}{2}\ln(2\pi) - \frac{1}{2}\gamma_k^\intercal \gamma_k\right\} \\
&+ \sum_k\left\{-\frac{D}{2}\ln(2\pi) + \frac{D}{2}\ln\tau_k + \ln|\Lambda_0| - \frac{\tau_k}{2}(\beta_k - m_0)^\intercal \Lambda_0(\beta_k - m_0)\right\} \\
&+ \sum_k\left\{(a_0 - 1)\ln\tau_k - b_0\tau_k\right\}
\end{aligned} \tag{8}
$$

We consider the following variational distribution used to approximate the posterior distribution of the parameters outlined previously.

$$
q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}\right) = q(\mathbf{Z})q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}) \tag{9}
$$

3.1 Coordinate Ascent Updates

As is standard in variational algorithms, we now seek the sequential updates of the factors in (9). For a particular variational parameter, the optimal distribution is found by taking the expectation of the joint distribution of all the random variables with respect to all of the *other* variational parameters, excluding the one of interest (**bishop:06**). Proceeding this way, we arrive at the following update equation for $q(\mathbf{Z})$,

$$
\ln q^*(\mathbf{Z}) = \mathbb{E}_{-q(\mathbf{Z})}\Big[\ln\left\{p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \boldsymbol{\gamma}\right)\right\}\Big] \tag{10}
$$

We adopt the convention that the expectation with respect to a negative subscript indicates an expectation taken with respect to the other variational parameters. Ignoring terms that are not functionally dependent on $\mathbf{Z}$, we can exponentiate both sides of (10) to obtain the optimal solution for $q(\mathbf{Z})$,

$$
q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{11}
$$

For the details in this calculation, refer to Appendix A. The form of this discrete distribution gives us $\mathbb{E}[z_{nk}] = r_{nk}$. If we consider the quantity $\ln\rho_{nk}$, defined in (12) and discussed in more detail in Appendix A, we note that the exact computation of $\ln\rho_{nk}$ involves four expectations taken with respect to the variational distribution $q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma})$.

$$
\begin{aligned}
\ln \rho_{nk} = &- \frac{1}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\tau})}[\ln \tau_k] - \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^{\intercal}\beta_k)^2] \\
&+ x_n^{\intercal}\mathbb{E}_{q(\boldsymbol{\gamma})}[\gamma_k] - \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln\left(\sum_j \exp\{x_n^{\intercal}\gamma_j\}\right)\right]
\end{aligned} \tag{12}
$$

We briefly consider each of these expectations. Since we used conjugate families, we know that $\mathbb{E}_{q(\boldsymbol{\tau})}[\ln \tau_k], \mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^{\intercal}\beta_k)^2]$, and $\mathbb{E}_{q(\boldsymbol{\gamma})}[\gamma_k]$ will have closed form expressions. The remaining expectation, $\mathbb{E}_{q(\boldsymbol{\gamma})}\big[\ln \sum_j \exp\{x_n^{\intercal}\gamma_j\}\big]$ presents a problem in that there lacks a closed form expression. Therefore, in order to complete this update, we use the following upper bound to approximate this quantity,

$$
\mathbb{E}_{q(\boldsymbol{\gamma})}\big[\ln \sum_j \exp\{x_n^{\intercal}\gamma_j\}\big] \approx \alpha_n + \varphi_n \tag{13}
$$

where $\varphi_n = \sum_j^K \frac{1}{2}\left(x_n^{\intercal}\mu_j - \alpha_n + \xi_{nj}\right) + \log(1 + e^{\xi_{nj}})$. The details for the approximation and how to find α_n and ξ_{nj} can be found in Appendix B. Using this in place of the problematic expectation in (12), we are able to obtain r_{nk} in closed form.

The remaining variational distribution can be found by considering the expectation of the joint density in (8) taken with respect to $q(\mathbf{Z})$, as derived above. The resulting variational distribution can be written up to constants as shown below,

$$
\begin{aligned}
\ln q^*(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma}) = &\sum_k \sum_n \mathbb{E}_{q(\mathbf{Z})}[z_{nk}] \ln \mathcal{N}\left(y_n \mid x_n^{\intercal}\beta_k, \tau_k^{-1}\right) + \sum_k \ln p\left(\beta_k, \tau_k\right) \\
&+ \sum_k \sum_n \mathbb{E}_{q(\mathbf{Z})}[z_{nk}] \left(x_n^{\intercal}\gamma_k - \ln \sum_{j=1}^K \exp\{x_n^{\intercal}\gamma_j\}\right) + \sum_k \ln \mathcal{N}(\gamma_k \mid 0, \mathrm{I}_D)
\end{aligned} \tag{14}
$$

Having just deriving the form for $q(\mathbf{Z})$, we can write out closed forms for the two expectation terms above. Noting in the summations above that the optimal distribution for $q(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma})$ has a sum involving only the γ_k's in the second line and a sum involving only the (β_k, τ_k)'s in the first line of (14), we can deduce $q(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma}) = \prod_k q(\beta_k, \tau_k) q(\gamma_k)$ and find the optimal distributions separately. With this factorization in mind, we obtain the following updates for the remaining variational parameters,

$$
q^*(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right) \tag{15}
$$

$$
q^*(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right) \tag{16}
$$

$$
q^*(\gamma_k) = \mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right) \tag{17}
$$

for $k = 1, \ldots, K$. The derivation and parameter definitions for (15) and (16) can be found in Appendix C, and the details for (17) can be found in Appendix B. Now that we have closed form updates for the variational distributions $q(\mathbf{Z})$ and $q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma})$, we can cycle through a two-step update procedure. In the first step (variational E-step), we compute the expectation of each of the z_{nk}'s using the variational distributions (15), (16), and (17). Then in the second step (variational M-step), we derive new optimal distributions using results from the E-step. We alternate between these two steps until the variational lower bound converges, as discussed in the section 3.2.

3.2 Evidence Lower Bound (ELBO)

In order to evaluate the convergence of the coordinate ascent algorithm, we can calculate the evidence lower bound using the updated variational parameters at the end of each iteration. Since the ELBO is monotonic increasing, we continue the coordinate ascent until the change in the ELBO between iterations falls below a predetermined tolerance. Note that the expectations taken below are with respect to the optimal variational distributions defined in the previous section.

$$
\begin{aligned}
\mathcal{L}(q) &= \sum_{z} \int\int\int q(\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z}) \ln \left\{ \frac{p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z})}{q(\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z})} \right\} d\boldsymbol{\beta} d\boldsymbol{\tau} d\boldsymbol{\gamma} \\
&= \mathbb{E}[\ln p\,(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z})] + \mathbb{E}[\ln p\,(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma})] + \mathbb{E}[\ln p\,(\boldsymbol{\gamma})] + \mathbb{E}[\ln p\,(\boldsymbol{\beta}, \boldsymbol{\tau})] \\
&\quad - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})] - \mathbb{E}[\ln q(\boldsymbol{\gamma})]
\end{aligned}
\tag{18}
$$

Details for each of the expectations can be found in Appendix D.

$$
\mathbb{E}[\ln p\,(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z})] = -\frac{1}{2} \sum_{n} \sum_{k} r_{nk} \Big\{ \ln(2\pi) - \psi(a_k) + \psi(b_k) + \frac{a_k}{b_k}(y_n - x_n^{\intercal} m_k)^2 + x_n^{\intercal} \mathrm{V}_k^{-1} x_n \Big\}
$$

$$
\mathbb{E}[\ln p\,(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma})] = \sum_{n} \sum_{k} r_{nk}\,(x_n^{\intercal} \mu_k - \alpha_n - \varphi_n)
$$

$$
\mathbb{E}[\ln p\,(\boldsymbol{\gamma})] = -\frac{K \cdot D}{2} \ln(2\pi) - \frac{1}{2} \sum_{k} \mu_k^{\intercal} \mu_k
$$

$$
\begin{aligned}
\mathbb{E}[\ln p\,(\boldsymbol{\beta}, \boldsymbol{\tau})] &= -K \left(\frac{D}{2} \ln(2\pi) - \ln|\Lambda_0| - a_0 \ln b_0 + \ln\Gamma(\alpha_0) \right) + \left(a_0 + \frac{D}{2} - 1 \right) \sum_{k} \psi(a_k) - \psi(b_k) \\
&\quad - \frac{1}{2} \sum_{k} \left\{ \frac{a_k}{b_k} \Big[(m_k - m_0)^{\intercal} \Lambda_0 (m_k - m_0) + b_0 \Big] + \mathrm{tr}\left(\Lambda_0 \mathrm{V}_k^{-1} \right) \right\}
\end{aligned}
$$

$$
\mathbb{E}[\ln q(\mathbf{Z})] = \sum_{n} \sum_{k} r_{nk} \ln r_{nk}
$$

$$
\begin{aligned}
\mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})] &= \sum_{k} \left(\frac{D}{2} + a_k - 1 \right) \big[\psi(a_k) - \psi(b_k) \big] + a_k(\ln b_k - 1) - \ln\Gamma(a_k) + \ln|\mathrm{V}_k| \\
&\quad - \frac{KD}{2} (\ln(2\pi) + 1)
\end{aligned}
$$

$$
\mathbb{E}[\ln q(\boldsymbol{\gamma})] = -\frac{KD}{2}(\ln(2\pi) + 1) + \sum_{k} \ln|\mathrm{Q}_k|
$$

4. Algorithm

Using the updates discussed in the previous section, we can formalize the variational algorithm below. Note that in the update of the r_{nk}'s, we use the approximation given in (13) to compute the update in closed form.

Algorithm 1 CAVI for Conditional Density Estimation

Result: An approximating distribution to the true posterior of $\boldsymbol{\theta}$

Input: Data $y_{1:N}, x_{1:N}$, number of components K, prior mean and precision for $\boldsymbol{\beta}_{1:K}$, prior shape, rate parameters for precision parameters $\tau_{1:K}$

Output: A variational density $q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}\right) = q(\mathbf{Z})q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}) = q(\mathbf{Z}) \prod\limits_k q(\beta_k, \tau_k)q(\gamma_k)$;

Initialize: Variational parameters $\mathbf{m}_{1:K}, \mathrm{V}_{1:K}, \boldsymbol{\mu}_{1:K}, \mathrm{Q}_{1:K}, a_{1:K}, b_{1:K}, \xi_{1:N,1:K}, \alpha_{1:N}$

while *the ELBO has not converged* **do**

  for $n \in \{1, \ldots, N\}$ **do**

    for $k \in \{1, \ldots, K\}$ **do**

      Set $r_{nk} \propto \exp\Big\{ - \frac{1}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}[\ln \tau_k] - \frac{1}{2}\mathbb{E}[\tau_k(y_n - x_n^\intercal \beta_k)^2] + x_n^\intercal \mathbb{E}[\gamma_k]$
      $-\mathbb{E}\Big[\ln\Big(\sum_{j=1}^K \exp\{x_n^\intercal \gamma_j\} \Big) \Big] \Big\}$

    end

  end

  for $n \in \{1, \ldots, N\}$ **do**

    for $k \in \{1, \ldots, K\}$ **do**

      Set $\xi_{nk} \leftarrow \sqrt{(x_n^\intercal \mu_k - \alpha_n)^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n}$

    end

    Set $\alpha_n \leftarrow \dfrac{\frac{1}{2}\left(\frac{K}{2} - 1\right) + \sum\limits_k \lambda(\xi_{nk})\mu_k^\intercal x_n}{\sum_k \lambda(\xi_{nk})}$

  end

  for $k \in \{1, \ldots, K\}$ **do**

    Set $\mathrm{Q}_k \leftarrow \mathrm{I}_D + 2\sum_n r_{nk}\lambda(\xi_{nk}) x_n x_n^\intercal$
    Set $\eta_k \leftarrow \sum_n r_{nk}\left[\frac{1}{2} + 2\lambda(\xi_{nk})\alpha_n\right] x_n$
    Set $\mu_k \leftarrow \mathrm{Q}_k^{-1}\eta_k$
    Set $\mathrm{V}_k \leftarrow \sum\limits_n r_{nk} x_n x_n^\intercal + \Lambda_0$
    Set $\zeta_k \leftarrow \sum_n r_{nk} y_n x_n + \Lambda_0 m_0$
    Set $m_k \leftarrow \mathrm{V}_k^{-1}\zeta_k$
    Set $a_k \leftarrow a_0 + N_k$
    Set $b_k \leftarrow b_0 + \frac{1}{2}[\sum\limits_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - \zeta_k^\intercal \mathrm{V}_k^{-1} \zeta_k]$

  end

  Compute ELBO using updated parameters

end

return $q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}\right)$

5. Alternative Bound

Instead of using the bound in (13), we can also consider the following bound which makes use of concavity of the log.

$$\log \sum_{k=1}^{K} e^{x_k} \leq \alpha \sum_{k=1}^{K} - \log \alpha - 1 \tag{19}$$

for any $x_k \in \mathbb{R}, \alpha \in \mathbb{R}$, where $x_k \sim q$. Taking expectation with respect to q, the right hand side is available in closed form (for most distributions) since $E_q[e^{x_k}]$ is nothing but the moment generating function of the distribution q. In the context of the Bayesian density regression problem, we can use bound in (19) to obtain the following bound

$$\begin{aligned} \mathbb{E}_{q(\boldsymbol{\gamma})}\Big[\ln \sum_j \exp\{x_n^\intercal \gamma_j\}\Big] &\leq \alpha_n \sum_j \mathbb{E}_{q(\boldsymbol{\gamma})}\Big[\exp\{x_n^\intercal \gamma_j\}\Big] - \log \alpha_n - 1 \\ &= \alpha_n \sum_j \exp\Big\{x_n^\intercal \mu_j + \frac{1}{2} x_n^\intercal \mathrm{Q}_j^{-1} x_n\Big\} - \log \alpha_n - 1 \end{aligned} \tag{20}$$

Choosing α_n to minimize the the right hand side of (20), we have

$$\alpha_n = \left(\sum_j \exp\Big\{x_n^\intercal \mu_j + \frac{1}{2} x_n^\intercal \mathrm{Q}_j^{-1} x_n\Big\}\right)^{-1}$$

Then, α_n can be updated as an additional variational to the coordinate ascent algorithm. Note that the only changes that we would have to make in the algorithm presented in section 4 are the expectation used to update r_{nk}, the updates for α_n in the second set of for loops, and mean/precision components of $q(\gamma_k)$.

6. Conclusion

In this project, we focused on one primary approximation to the difficult-to-compute expectation of the log sum of exponentials, $\mathbb{E}_{q(\boldsymbol{\gamma})}\big[\ln \sum_j \exp\{x_n^\intercal \gamma_j\}\big]$. Alternatively, we also considered a simpler bound in section 5, which only requires us to compute moment generating functions. As has been noted in previous work and simulations, the former bound provides better approximations when the variance of $q(\boldsymbol{\gamma})$ is extremely large because the approximation is asymptotically optimal (**bouchard:07**), but in most other cases, the performance deteriorated (**Depraetere:17**). Although both of these methods require additional variational parameters, this increase in model complexity is offset by the faster inference that variational Bayesian methods provide. As a result, even though each iteration involves $2K$ $(D \times D)$-matrix inversions to update the variational distributions for β_k and γ_k, we expect the number of iterations in the coordinate ascent algorithm to be far fewer than if we use a traditional Gibbs sampling scheme.

While we only considered two candidate approximations in this project, there are numerous ways to approach the problem. Although in a slightly different context, Depraeteret and

Vandebroek (2017) provide several approximations to the same quantity of interest, all of which admit closed form updates so that variational Bayesian methods are tractable.

Appendix A. Variational Update for $q(\mathbf{Z})$

Taking the expectation with respect to the other variational parameters, we can derive the following variational distribution for $\mathbf{Z}$,

$$\begin{aligned}\ln q^*(\mathbf{Z}) = \sum_n \sum_k z_{nk} \Bigg\{ & -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\tau})}[\ln \tau_k] - \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^\intercal \beta_k)^2] \\ & + x_n^\intercal \mathbb{E}_{q(\boldsymbol{\gamma})}[\gamma_k] - \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln \sum_j \exp\{x_n^\intercal \gamma_j\}\right]\Bigg\} \\ = \sum_n \sum_k z_{nk} \ln \rho_{nk} &\end{aligned}$$

where we have defined

$$\begin{aligned}\ln \rho_{nk} = & -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\tau})}[\ln \tau_k] - \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^\intercal \beta_k)^2] \\ & + x_n^\intercal \mathbb{E}_{q(\boldsymbol{\gamma})}[\gamma_k] - \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln \sum_j \exp\{x_n^\intercal \gamma_j\}\right]\end{aligned} \tag{21}$$

Exponentiating and normalizing, we have

$$q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{22}$$

For the discrete distribution $q^*(\mathbf{Z})$ given in (22) above, we have $E[z_{nk}] = r_{nk}$. Note, however, that in order to compute the expectation in closed form, we need an expression for the four expectations involved in the quantity $\ln \rho_{nk}$, as defined in (21).

From the results derived in Appendix C, we know that $q^*(\tau_k) = \mathrm{Ga}(\tau_k \mid a_k, b_k)$. We can then compute the following expectation with respect to $q^*(\boldsymbol{\tau})$).

$$E_{q(\boldsymbol{\tau})}[\ln \tau_k] = \psi(a_k) - \psi(b_k) \tag{23}$$

Again from Appendix C, we can then compute the following expectation with respect to $q^*(\beta_k, \tau_k)$.

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}\left[\tau_k(y_n - x_n^\intercal \beta_k)^2\right] &= \mathbb{E}\left[\tau_k \left(y_n + m_k^\intercal x_n x_n^\intercal m_k + \mathrm{tr}\left(x_n x_n^\intercal \left(\tau_k \mathrm{V}_k\right)^{-1}\right) - 2 y_n x_n^\intercal m_k\right)\right] \\ &= \frac{a_k}{b_k}\left(y_n^2 + m_k^\intercal x_n x_n^\intercal m_k - 2 y_n x_n^\intercal m_k\right) + \mathrm{tr}\left(x_n x_n^\intercal \mathrm{V}_k^{-1}\right) \\ &= \frac{a_k}{b_k}(y_n - m_k^\intercal x_n)^2 + x_n^\intercal \mathrm{V}_k^{-1} x_n\end{aligned} \tag{24}$$

From the expression derived in (29) of Appendix B, we have $q^*(\gamma_k) = \mathcal{N}(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1})$, then we have

$$\mathbb{E}_{q(\gamma_k)}[\gamma_k] = \mu_k \tag{25}$$

Using the bound discussed in Appendix B, equation (28), we can then compute the following expectation with respect to $q^*(\boldsymbol{\gamma})$.

$$
\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln\sum_{j}^{K}\exp\{x_n^\intercal\gamma_j\}\right] \\
&\approx \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\alpha_n+\sum_{j=1}^{K}\frac{x_n^\intercal\gamma_j-\alpha_n-\xi_{nj}}{2}+\lambda(\xi_{nj})\left((x_n^\intercal\gamma_j-\alpha_n)^2-\xi_{nj}^2\right)+\log\left(1+e^{\xi_{nj}}\right)\right] \\
&=\alpha_n+\sum_{j}^{K}\frac{1}{2}\left(x_n^\intercal\mu_j-\alpha_n-\xi_{nj}\right)+\lambda(\xi_{nj})\left((x_n^\intercal\mu_j-\alpha_n)^2-\xi_{nj}^2+x_n^\intercal\mathrm{Q}_j^{-1}x_n\right)+\log(1+e^{\xi_{nj}}) \\
&=\alpha_n+\sum_{j}^{K}\frac{1}{2}\left(x_n^\intercal\mu_j-\alpha_n-\xi_{nj}\right)+\log(1+e^{\xi_{nj}})
\end{aligned}
\tag{26}
$$

since $(x_n^\intercal\mu_j-\alpha_n)^2-\xi_{nj}^2=-x_n^\intercal\mathrm{Q}_j^{-1}x_n$. Gathering the results in (23), (24), (25), and (26), and substituting these into (21), we can compute $\mathbb{E}[z_{nk}]=r_{nk}$ in closed form.

Appendix B. Variational Updates for $q(\gamma_k)$

For the variational distribution for $\gamma_k, k=1,\ldots,K$, we first note the following bound (**bouchard:07**), $\sum_{j=1}^{K}e^{t_j}\leq\prod_{j=1}^{K}(1+e^{t_j})$. Setting $t_j=x_n^\intercal\gamma_j-\alpha_n$ and then taking log, we have the following bound:

$$
\log\left(\sum_{j=1}^{K}\exp\{x_n^\intercal\gamma_j\}\right)\leq\alpha_n+\sum_{j=1}^{K}\log\left(1+\exp\{x_n^\intercal\gamma_j-\alpha_n\}\right)
\tag{27}
$$

We can further bound this by using the following tangential bound (**jj:2001**),

$$
\log(1+e^x)\leq\frac{x-\xi}{2}+\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)(x^2-\xi^2)+\log\left(1+e^{\xi}\right)
$$

then we arrive at the following bound:

$$
\log\left(\sum_{j=1}^{K}\exp\{x_n^\intercal\gamma_j\}\right)\leq\alpha_n+\sum_{j=1}^{K}\frac{x_n^\intercal\gamma_j-\alpha_n-\xi_{nj}}{2}+\lambda(\xi_{nj})\left((x_n^\intercal\gamma_j-\alpha_n)^2-\xi_{nj}^2\right)+\log\left(1+e^{\xi_{nj}}\right)
\tag{28}
$$

where $\lambda(\xi)=\frac{1}{4\xi}\tanh\left(\frac{\xi}{2}\right)$. Then we can substitute this back into $\ln q^*(\gamma_k)$ to obtain an approximation for the left hand side of (27), thus allowing us to obtain a closed form for the variational distribution. Note that all of the equalities above are written up to constants.

$$
\begin{aligned}
\ln q^*(\gamma_k) &= -\frac{1}{2}\gamma_k^\intercal\gamma_k + \sum_n r_{nk}x_n^\intercal\gamma_k - \sum_n r_{nk}\ln\left(\sum_j \exp\{x_n^\intercal\gamma_j\}\right) \\
&\approx -\frac{1}{2}\gamma_k^\intercal\gamma_k + \gamma_k^\intercal\sum_n r_{nk}x_n \\
&\quad - \sum_n r_{nk}\left\{\alpha_n + \sum_{j=1}^K \frac{x_n^\intercal\gamma_j - \alpha_n - \xi_{nj}}{2} + \lambda(\xi_{nj})\left((x_n^\intercal\gamma_j-\alpha_n)^2 - \xi_{nj}^2\right) + \log\left(1+e^{\xi_{nj}}\right)\right\} \\
&= -\frac{1}{2}\gamma_k^\intercal\gamma_k + \gamma_k^\intercal\sum_n r_{nk}x_n - \sum_n r_{nk}\left\{\frac{1}{2}\gamma_k^\intercal x_n + \lambda\left(\xi_{nj}\right)\left(\gamma_k^\intercal x_n x_n^\intercal\gamma_k - 2\alpha_n\gamma_k^\intercal x_n\right)\right\} \\
&= -\frac{1}{2}\gamma_k^\intercal\left(\mathrm{I}_D + 2\sum_n r_{nk}\lambda(\xi_{nk})x_n x_n^\intercal\right)\gamma_k + \gamma_k'\left(\sum_n r_{nk}\left(\frac{1}{2} + 2\lambda\left(\xi_{nk}\right)\alpha_n x_n\right)\right)
\end{aligned}
$$

Exponentiating, we can recover $q^*(\gamma_k) = \mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right)$, where

$$
\begin{aligned}
\mu_k &= \mathrm{Q}_k^{-1}\eta_k \\
\eta_k &= \sum_n r_{nk}\left(\frac{1}{2} + 2\lambda(\xi_{nj})\alpha_n\right)x_n \\
\mathrm{Q}_k &= \mathrm{I}_D + 2\sum_n r_{nk}\lambda(\xi_{nk})x_n x_n^\intercal
\end{aligned}
\tag{29}
$$

The additional parameters introduced in the two upper bounds can be updated using the following equations (**Depraetere:17**),

$$
\xi_{nk} = \sqrt{\left(\mu_k^\intercal x_n - \alpha_n\right)^2 + x_n^\intercal\mathrm{Q}_k^{-1}x_n} \qquad \forall k, n
$$

$$
\alpha_n = \frac{\frac{1}{2}\left(\frac{K}{2}-1\right) + \sum_{j=1}^K \lambda\left(\xi_{nj}\right)\mu_j^\intercal x_n}{\sum_{j=1}^K \lambda\left(\xi_{nj}\right)} \qquad \forall n
$$

Appendix C. Variational Updates for $q(\beta_k, \tau_k)$

Using results from Appendix A, we can write the following expression for the joint variational distribution of (β_k, τ_k),

$$
\begin{aligned}
\ln q^*(\beta_k, \tau_k) = &\sum_n -\frac{1}{2}r_{nk}\tau_k\left(y_n^2 + \beta_k^\intercal x_n x_n^\intercal\beta_k - 2y_n\beta_k^\intercal x_n\right) + \frac{r_{nk}}{2}\ln\tau_k + \frac{D}{2}\ln\tau_k \\
&- \frac{\tau_k}{2}\left(\beta_k^\intercal\Lambda_0\beta_k + m_0^\intercal\Lambda_0 m_0 - 2\beta_k^\intercal\Lambda_0 m_0\right) + (a_0-1)\ln\tau_k - b_0\tau_k
\end{aligned}
\tag{30}
$$

We first consider terms on the right hand side of (30) that depend on β_k to find $\ln q^\star(\beta_k \,|\, \tau_k)$, giving

$$
\ln q^\star(\beta_k \mid \tau_k) = -\frac{\tau_k}{2}\beta_k^\intercal\Big[\sum_n r_{nk}x_n x_n^\intercal + \Lambda_0\Big]\beta_k + \tau_k\beta_k^\intercal\Big[\sum_n r_{nk}y_n x_n + \Lambda_0 m_0\Big] \tag{31}
$$

$$q^*(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right) \tag{32}$$

$$\begin{aligned} \mathrm{V}_k &= \sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 \\ \zeta_k &= \sum_n r_{nk} y_n x_n + \Lambda_0 m_0 \\ m_k &= \mathrm{V}_k^{-1} \zeta_k \end{aligned} \tag{33}$$

Then we can make use of the relation $\ln q^*(\tau_k) = \ln q^\star(\beta_k, \tau_k) - \ln q^\star(\beta_k \mid \tau_k)$, where the quantities on the right hand side come from (30) and (32). Note that equality below is written up to constants, keeping only terms involving τ_k.

$$\begin{aligned} \ln q^*(\tau_k) = (a_0 + N_k - 1)\ln \tau_k - \tau_k \Bigg\{ & b_0 + \frac{1}{2}\left(\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - m_k^\intercal \mathrm{V}_k m_k\right) \\ & + \frac{1}{2}\beta_k^\intercal \left(\sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 - \mathrm{V}_k\right)\beta_k \\ & - 2\beta_k^\intercal \left(\sum_n r_{nk} y_n x_n + \Lambda_0 m_0 - \mathrm{V}_k m_k\right)\Bigg\} \end{aligned} \tag{34}$$

Exponentiating, we arrive at the following distribution

$$q^*(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right) \tag{35}$$

where we have defined

$$\begin{aligned} a_k &= a_0 + \frac{N_k}{2} \\ b_k &= b_0 + \frac{1}{2}\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - \zeta_k^\intercal \mathrm{V}_k^{-1} \zeta_k \end{aligned} \tag{36}$$

The expression for b_k arises by noting that the three following simplifications for the summation terms in the coefficient of τ_k in (34),

$$\begin{aligned} &\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - m_k^\intercal \mathrm{V}_k m_k = \sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - \zeta_k^\intercal \mathrm{V}_k^{-1} \zeta_k \\ &\sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 - \mathrm{V}_k = 0 \\ &\sum_n r_{nk} y_n x_n + \Lambda_0 m_0 - \mathrm{V}_k m_k = 0 \end{aligned}$$

where the first equality holds by expanding $m_k^\intercal \mathrm{V}_k m_k = \zeta_k^\intercal \left(\mathrm{V}_k^{-1}\right)^\intercal \mathrm{V}_k \mathrm{V}_k^{-1} \zeta_k = b_k^\intercal \mathrm{V}_k^{-1} \zeta_k$. The second quality holds by recalling the definition of V_k in (33), and the third equality holds by observing from (33) that $\mathrm{V}_k m_k = \zeta_k$

Appendix D. Variational Lower Bound

As seen in (18) of section 3.2, we need to calculate seven expectations (taken with respect to the variational distribution $q(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma})$. Below, we compute each of these expectations in detail, making extensive use of the variational distributions derived in Appendix A, B, and C.

$$\begin{aligned}
\mathbb{E}[\ln p\,(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z})] &= \sum_n \sum_k \mathbb{E}[z_{nk}]\mathbb{E}\left[\ln \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right)\right] \\
&= -\frac{1}{2}\sum_n \sum_k r_{nk}\Big\{\ln(2\pi) - \mathbb{E}[\ln \tau_k] + \mathbb{E}\Big[\tau_k\left(y_n - x_n^\intercal \beta_k\right)^2\Big]\Big\} \\
&= -\frac{1}{2}\sum_n \sum_k r_{nk}\Big\{\ln(2\pi) - (\psi(a_k) - \psi(b_k)) + x_n^\intercal \mathrm{V}_k^{-1} x_n \\
&\qquad + \frac{a_k}{b_k}(y_n - x_n^\intercal m_k)^2\Big\}
\end{aligned} \tag{37}$$

$$\begin{aligned}
\mathbb{E}[\ln p\,(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma})] &= \sum_n \sum_k \mathbb{E}[z_{nk}]\left(\mathbb{E}[x_n^\intercal \gamma_k] - \mathbb{E}\Big[\ln \sum_j \exp\{x_n^\intercal \gamma_j\}\Big]\right) \\
&\approx \sum_n \sum_k r_{nk}\left(x_n^\intercal \mu_k - \alpha_n - \varphi_n\right)
\end{aligned} \tag{38}$$

where $\varphi_n = \sum_{j=1}^{K} \frac{1}{2}\left(x_n^\intercal \mu_j - \alpha_n - \xi_{nj}\right) + \log(1 + e^{\xi_{nj}})$. Here, we make use of the result in (26), where we take the expectation of the upper bound previously derived.

$$\begin{aligned}
\mathbb{E}[\ln p\,(\boldsymbol{\gamma})] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right)\Big] \\
&= -\frac{K \cdot D}{2}\ln(2\pi) - \frac{1}{2}\sum_k \mu_k^\intercal \mu_k + \mathrm{tr}\left(\mathrm{Q}_k^{-1}\right)
\end{aligned} \tag{39}$$

$$\begin{aligned}
\mathbb{E}[\ln p\,(\boldsymbol{\beta}, \boldsymbol{\tau})] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\beta_k \mid m_0, (\tau_k \Lambda_0)^{-1}\right)\Big] + \mathbb{E}\Big[\ln \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right)\Big] \\
&= \left(a_0 + \frac{D}{2} - 1\right)\sum_k \psi(a_k) - \psi(b_k) \\
&\quad - \frac{1}{2}\sum_k \left\{\frac{a_k}{b_k}\Big[(m_k - m_0)^\intercal \Lambda_0 (m_k - m_0) + b_0\Big] + \mathrm{tr}\left(\Lambda_0 \mathrm{V}_k^{-1}\right)\right\} \\
&\quad - K\left(\frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\Lambda_0| - a_0 \ln b_0 + \ln \Gamma(\alpha_0)\right)
\end{aligned} \tag{40}$$

where we make use of $\mathbb{E}[\tau_k(\beta_k - m_0)\Lambda_0(\beta_k - m_0)^\intercal] = \frac{a_k}{b_k}(m_k - m_0)^\intercal \Lambda_0 (m_k - m_0) + \mathrm{tr}\left(\Lambda_0 \mathrm{V}_k^{-1}\right)$. The other expectations can be calculated using results derived in Appendix A. In the following expectation, we make use of the established result that $E[z_{nk}] = r_{nk}$.

$$\mathbb{E}[\ln q(\mathbf{Z})] = \sum_n \sum_k r_{nk} \ln r_{nk} \tag{41}$$

$$\begin{aligned}\mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\beta_k \mid m_k, (\tau_k V_k)^{-1}\right)\Big] + \mathbb{E}\Big[\ln \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right)\Big] \\ &= \sum_k \left(\frac{D}{2} + a_k - 1\right)\big[\psi(a_k) - \psi(b_k)\big] + a_k \ln b_k - a_k - \ln\Gamma(a_k) + \ln|\mathrm{V}_k| \\ &\qquad - \frac{KD}{2}\left(\ln(2\pi) + 1\right)\end{aligned} \tag{42}$$

where we make use of $\mathbb{E}[\tau_k(\beta_k - m_k)^\intercal \mathrm{V}_k(\beta_k - m_k)] = D$. The other expectations can be calculated using results derived in Appendix A.

$$\begin{aligned}\mathbb{E}[\ln q(\boldsymbol{\gamma})] &= \sum_k \mathbb{E}\Big[\mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right)\Big] \\ &= -\frac{KD}{2}(\ln(2\pi) + 1) + \frac{1}{2}\sum_k \ln|\mathrm{Q}_k|\end{aligned} \tag{43}$$

since $\mathbb{E}[(\gamma_k - \mu_k)^\intercal \mathrm{Q}_k(\gamma_k - \mu_k)] = D$.

Appendix E. Variable Selection

Recall the previous setup, where we denote $\mathbf{y} = \{y_1, \ldots, y_N\}, \mathbf{X} = \{x_1, \ldots, x_N\}$, where each $x_n \in \mathbb{R}^D$. Let $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}$ and $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_k\}$ denote the D-dimensional coefficient vectors in the mixture weights and in gaussian mixture components, respectively. Finally, let $\boldsymbol{\tau} = \{\tau_1, \ldots, \tau_k\}$ denote the precision parameters for each of the gaussian components in the mixture density. The mixture density can then be written explicitly as

$$p\left(y_n \mid x_n, \boldsymbol{\beta}, \boldsymbol{\tau}\right) = \sum_k \pi_k(x_n) \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right), \quad \pi_k(x_n) = \frac{\exp\{x_n^\intercal \gamma_k\}}{\sum\limits_{j=1}^{K} \exp\{x_n^\intercal \gamma_j\}} \tag{44}$$

Conditioning on the same set of latent variables, $\mathbf{Z} = \{z_1, \ldots, z_N\}$, used indicate the membership for each of the response variables, we obtain the mixture density in product form.

$$p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) = \prod_n \prod_k \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right)^{z_{nk}} \tag{45}$$

$$p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) = \prod_n \prod_k \pi_k(x_n)^{z_{nk}} = \prod_n \prod_k \left(\frac{\exp\{x_n^\intercal \gamma_k\}}{\sum_{j=1}^{K} \exp\{x_n^\intercal \gamma_j\}}\right)^{z_{nk}} \tag{46}$$

For $\boldsymbol{\gamma}$, we consider an independent standard normal prior on the γ_k's, given by

$$p(\boldsymbol{\gamma}) = \prod_k p(\gamma_k) = \prod_k \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right) \tag{47}$$

Instead of the normal-gamma prior that we used in the standard setup, we place a gamma prior on τ_k's:

$$p(\boldsymbol{\tau}) = \prod_k p(\tau_k) = \prod_k \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right) \tag{48}$$

Finally, to incorporate variable selection we consider the *rows* of $\boldsymbol{\beta}$, as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} | & | & & | \\ \beta_1 & \beta_2 & \ldots & \beta_K \\ | & | & & | \end{bmatrix} = \begin{bmatrix} — & \tilde{\beta}_1 & — \\ & \vdots & \\ — & \tilde{\beta}_D & — \end{bmatrix}$$

Using the above formulation, $\tilde{\beta}_d = (\beta_{1d}, \beta_{2d}, \ldots, \beta_{Kd})^\intercal \in \mathbb{R}^K$ represents the d-th row of the coefficient matrix. In other words, $\tilde{\beta}_d$ represents the coefficient vector for the d-th predictor across all K clusters. We place independent spike and slab priors on each *row* vectors of $\boldsymbol{\beta}$,

$$\tilde{\beta}_d \sim \pi \mathcal{N}\left(\tilde{\beta}_d \mid 0, \xi_0^{-1} \cdot \mathrm{I}_K\right) + (1 - \pi)\delta_0(\tilde{\beta}_d) \tag{49}$$

for $\tilde{\beta}_d$ for $d = 1, \ldots, D$.

Introducing $\boldsymbol{\omega} = \{\omega_1, \ldots, \omega_D\}$, where $\omega_d \sim \mathrm{Ber}(\omega_d \mid \pi)$, we can rewrite the prior on $\boldsymbol{\beta}$ as:

$$p\left(\boldsymbol{\beta}\right) = \prod_d p\left(\tilde{\beta}_d, \omega_d\right) = \prod_d \mathcal{N}\left(\tilde{\beta}_d \mid 0, \xi_0^{-1} \cdot \mathrm{I}_K\right)^{\omega_d} \pi^{\omega_d}(1 - \pi)^{1 - \omega_d} \tag{50}$$

E.1 Approximating Distribution

The variational parameters of interest are $\boldsymbol{\theta} = (\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega})$. We consider an approximating distribution of the following form

$$q(\boldsymbol{\theta}) = q\left(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}\right) = q(\mathbf{Z}) \cdot q(\boldsymbol{\tau}, \boldsymbol{\gamma}) \cdot \prod_d q(\tilde{\beta}_d, \omega_d) \tag{51}$$

It can be shown q as formulated above in (51), further factorizes into

$$q\left(\mathbf{Z}, \boldsymbol{\tau}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\omega}\right) = q(\mathbf{Z}) \cdot q(\boldsymbol{\tau}) \cdot q(\boldsymbol{\gamma}) \cdot \prod_d q(\tilde{\beta}_d, \omega_d) \tag{52}$$

where we do not assume anything further about the functional form of each of the factorized approximating distributions.

E.2 Joint Likelihood

In order to find closed form updates for the coorindate ascent algorithm, we first compute the log of the joint distribution of these random variables

$$\begin{aligned}
\ln p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) &= \ln \left\{ p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \boldsymbol{\gamma}\right) p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) p\left(\tilde{\beta}_d, \omega_d\right) p\left(\tau\right) \right\} \\
&= \sum_n \sum_k z_{nk} \left\{ -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\tau_k - \frac{\tau_k}{2}\left(y_n - x_n^{\intercal}\beta_k\right)^2 \right\} \\
&+ \sum_n \sum_k z_{nk} \left\{ x_n^{\intercal}\gamma_k - \ln \sum_{j=1}^{K} \exp\{x_n^{\intercal}\gamma_j\} \right\} + \sum_k \left\{ -\frac{D}{2}\ln(2\pi) - \frac{1}{2}\gamma_k^{\intercal}\gamma_k \right\} \\
&+ \sum_d \omega_d \left\{ -\frac{K}{2}\ln(2\pi) + \frac{K}{2}\ln\xi_0 - \frac{\xi_0}{2}\tilde{\beta}_d^{\intercal}\tilde{\beta}_d \right\} + \omega_d \ln\pi + (1-\omega_d)\ln(1-\pi) \\
&+ \sum_k \left\{ (a_0 - 1)\ln\tau_k - b_0\tau_k \right\}
\end{aligned} \tag{53}$$

E.3 CAVI Updates

Proceeding in standard fashion by taking the expectation with respect to the variational distribution, we can derive the functional form for each of the approximating distributions and their corresponding coordinate ascent updates.

For $q\left(\tilde{\beta}_d \mid \omega_d = 1\right)$, it is helpful to rewrite to first term in the log joint likelihood (mixsture density) in terms of $\tilde{\beta}_d$. Then, we have the following variational distribution with equality written up to constants

$$\ln q^{\star}\left(\tilde{\beta}_d \mid \omega_d = 1\right) = -\frac{1}{2}\tilde{\beta}_d^{\intercal}\mathrm{U}_d\tilde{\beta}_d + \tilde{\beta}_d^{\intercal}\eta_d - \frac{\xi_0}{2}\tilde{\beta}_d^{\intercal}\tilde{\beta}_d \tag{54}$$

$$= -\frac{1}{2}\tilde{\beta}_d^{\intercal}\left[\mathrm{U}_d + \xi_0\mathrm{I}_K\right]\tilde{\beta}_d + \tilde{\beta}_d^{\intercal}\eta_d \tag{55}$$

Note that we have used the following result (equality written up to constants in $\tilde{\beta}_d$).

$$E_{-q(\boldsymbol{\beta})}\left[\frac{1}{2}\sum_n\sum_k -z_{nk}\tau_k\left(y_n - x_n^\intercal\beta_k\right)^2\right] = -\frac{1}{2}\sum_d \tilde{\beta}_d^\intercal \mathrm{U}_d\tilde{\beta}_d + \sum_d \tilde{\beta}_d^\intercal \eta_d \tag{56}$$

From (55), we have

$$q^\star(\tilde{\beta}_d) = \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right)$$

where $m_d = \mathrm{Q}_d^{-1}\eta_d, \mathrm{Q}_d = \mathrm{U}_d + \xi_0\mathrm{I}_K$, and $\eta_d = \zeta_d - \frac{1}{2}\sum_{j\neq d}\mathrm{R}_{dj}\omega_j m_j$. We have made use of the following matrices

$$\mathrm{U}_d = \begin{bmatrix} \frac{a_1}{b_1}\sum_n r_{n1}x_{nd}^2 & & \\ & \ddots & \\ & & \frac{a_K}{b_K}\sum_n r_{nK}x_{nd}^2 \end{bmatrix} \quad \mathrm{R}_{dj} = \begin{bmatrix} \frac{a_1}{b_1}\sum_n r_{n1}x_{nd}x_{nj} & & \\ & \ddots & \\ & & \frac{a_K}{b_K}\sum_n r_{nK}x_{nd}x_{nj} \end{bmatrix}$$

$$\zeta_d = \begin{bmatrix} \frac{a_1}{b_1}\sum_n r_{n1}x_{nd}y_n \\ \vdots \\ \frac{a_K}{b_K}\sum_n r_{nK}x_{nd}y_n \end{bmatrix}$$

Note that we restricted the approximating distribution for $p\left(\boldsymbol{\beta}, \boldsymbol{\omega}\right)$ to be of the form

$$q\left(\boldsymbol{\beta}, \boldsymbol{\omega}\right) = \prod_d q\left(\tilde{\beta}_d, \omega_d\right)$$

where each of the d factors then inherits a spike and slab density of the form

$$q\left(\tilde{\beta}_d, \omega_d\right) = \begin{cases} \lambda_d\mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right), & \text{if } \omega_d = 1 \\ (1-\lambda_d)\delta_0\left(\tilde{\beta}_d\right) & \text{if } \omega_d = 0 \end{cases}$$

The variational parameters $(m_d, \mathrm{Q}_d, \lambda_d)$ can be updated using the following by differentiating the variational lower bound with respect to each of the parameters, setting the resulting partial derivatives to zero and solving for each parameter. Doing so yields following:

$$\mathrm{Cov}(\tilde{\beta}_d \mid \omega_d = 1) \approx \mathrm{Q}_d^{-1} \quad = (\mathrm{U}_d + \xi_0\mathrm{I}_K)^{-1} \tag{57}$$

$$\mathbb{E}[\tilde{\beta}_d \mid \omega_d = 1] \approx m_d \quad = \mathrm{Q}_d^{-1}\eta_d \tag{58}$$

$$\frac{p\left(\omega_d = 1 \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}\right)}{p\left(\omega_d = 0 \mid \mathbf{X}, \mathbf{y}, \boldsymbol{\theta}\right)} \approx \frac{\lambda_d}{1-\lambda_d} = \frac{\pi}{1-\pi}\cdot\log\xi_0^{\frac{K}{2}}\cdot\exp\left\{\frac{1}{2}m_d^\intercal\eta_d\right\} \tag{59}$$

E.4 Commonly Used Expectations

Two useful formulations of the approximating density: $q\left(\tilde{\beta}_d, \omega_d\right)$

$$q(\tilde{\beta}_d \mid w_d) = \omega_d \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) + (1-\omega_d)\delta_0(\tilde{\beta}_d) \tag{60}$$

$$q(\tilde{\beta}_d) = \lambda_d \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) + (1-\lambda_d)\delta_0(\tilde{\beta}_d) \tag{61}$$

$$\begin{aligned}
\mathbb{E}\left[\tilde{\beta}_d\right] &= \lambda_d m_d \\
\mathbb{E}\left[\tilde{\beta}_d \mid w_d\right] &= w_d m_d \\
\mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal \mid w_d\right] &= w_d \left(\mathrm{Q}_d^{-1} + m_d m_d^\intercal\right) \\
\mathrm{Var}\left(\tilde{\beta}_d \mid w_d\right) &= w_d \mathrm{Q}_d^{-1} + w_d(1-w_d) m_d m_d^\intercal \\
\mathbb{E}\left[\tilde{\beta}_d - m_d \mid w_d\right] &= (w_d - 1) m_d \\
\mathbb{E}\left[\tilde{\beta}_d^\intercal \mathrm{U}_d \tilde{\beta}_d\right] &= \lambda_d \mathrm{tr}\left(\mathrm{U}_d \mathrm{Q}_d^{-1}\right) + \lambda_d m_d^\intercal \mathrm{U}_d m_d \\
\mathbb{E}\left[(\tilde{\beta}_d - m_d)^\intercal \mathrm{Q}_d (\tilde{\beta}_d - m_d) \mid w_d\right] &= w_d \cdot K \\
\mathbb{E}\left[w_d (\tilde{\beta}_d - m_d)^\intercal \mathrm{Q}_d (\tilde{\beta}_d - m_d)\right] &= K \cdot \lambda_d \\
\mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal\right] &= \lambda_d \left(\mathrm{Q}_d^{-1} + m_d m_d^\intercal\right) \\
\mathrm{Var}\left(\tilde{\beta}_d\right) &= \lambda_d \mathrm{Q}_d^{-1} + \lambda_d(1-\lambda_d) m_d m_d^\intercal
\end{aligned}$$

E.5 Expectation Calculations Details

$$
\begin{aligned}
\mathbb{E}\left[\tilde{\beta}_d\right] &= \int \tilde{\beta}_d \cdot \lambda_d \cdot \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) d\beta_d = \lambda_d m_d \\
\mathbb{E}\left[\tilde{\beta}_d \mid w_d\right] &= \int \tilde{\beta}_d \cdot w_d \cdot \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) d\tilde{\beta}_d = w_d m_d \\
\mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal \mid w_d\right] &= \int \tilde{\beta}_d \tilde{\beta}_d^\intercal \cdot w_d \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) d\tilde{\beta}_d = w_d\left(\mathrm{Q}_d^{-1} + m_d m_d^\intercal\right) \\
\mathrm{Var}\left(\tilde{\beta}_d \mid w_d\right) &= \mathbb{E}\left[(\tilde{\beta}_d - \mathbb{E}(\tilde{\beta}_d))(\tilde{\beta}_d - \mathbb{E}(\tilde{\beta}_d))^\intercal \mid w_d\right] \\
&= \mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal \mid w_d\right] + \mathbb{E}\left[\tilde{\beta}_d \mid w_d\right]\mathbb{E}\left[\tilde{\beta}_d \mid w_d\right]^\intercal - 2\mathbb{E}\left[\tilde{\beta}_d \mid w_d\right]\mathbb{E}\left[\tilde{\beta}_d \mid w_d\right]^\intercal \\
&= w_d\left(m_d m_d^\intercal + \mathrm{Q}_d^{-1}\right) + w_d^2 m_d m_d^\intercal - 2 w_d^2 m_d m_d^\intercal \\
&= w_d \mathrm{Q}_d^{-1} + w_d(1 - w_d) m_d m_d^\intercal \\
\mathbb{E}\left[\tilde{\beta}_d - m_d \mid w_d\right] &= w_d m_d - m_d = (w_d - 1) m_d \\
\mathbb{E}\left[\tilde{\beta}_d^\intercal \mathrm{U}_d \tilde{\beta}_d\right] &= \mathbb{E}(\tilde{\beta}_d))^\intercal \mathrm{U}_d \mathbb{E}(\tilde{\beta}_d)) + \mathrm{tr}\left(\mathrm{U}_d \mathrm{Var}\left(\tilde{\beta}_d\right)\right) \\
&= \lambda_d^2 m_d^\intercal \mathrm{U}_d m_d + \mathrm{tr}\left(\mathrm{U}_d\left(\lambda_d \mathrm{Q}_d^{-1} + \lambda_d(1 - \lambda_d) m_d m_d^\intercal\right)\right. \\
&= \lambda_d^2 m_d^\intercal \mathrm{U}_d m_d + \lambda_d \mathrm{tr}\left(\mathrm{U}_d \mathrm{Q}_d^{-1}\right) + \lambda_d(1 - \lambda_d) m_d^\intercal \mathrm{U}_d m_d \\
&= \lambda_d \mathrm{tr}\left(\mathrm{U}_d \mathrm{Q}_d^{-1}\right) + \lambda_d m_d^\intercal \mathrm{U}_d m_d \\
\mathbb{E}\left[(\tilde{\beta}_d - m_d)^\intercal \mathrm{Q}_d (\tilde{\beta}_d - m_d) \mid w_d\right] &= \mathbb{E}\left[(\tilde{\beta}_d - m_d)^\intercal \mid w_d\right]^\intercal \mathrm{Q}_d \mathbb{E}\left[(\tilde{\beta}_d - m_d)^\intercal \mid w_d\right] + \mathrm{tr}\left(\mathrm{Q}_d \mathrm{Var}\left(\tilde{\beta}_d \mid w_d\right)\right) \\
&= (w_d - 1)^2 m_d^\intercal \mathrm{Q}_d m_d + w_d \mathrm{tr}\left(\mathrm{Q}_d \mathrm{Q}_d^{-1}\right) + w_d(1 - w_d) m_d^\intercal \mathrm{Q}_d m_d \\
&= (w_d^2 - 2w_d + 1) m_d^\intercal \mathrm{Q}_d m_d + w_d \mathrm{tr}\left(\mathrm{I}_K\right) + w_d m_d^\intercal \mathrm{Q}_d m_d - w_d^2 m_d^\intercal \mathrm{Q}_d m_d \\
&= K \cdot w_d + m_d^\intercal \mathrm{Q}_d m_d - w_d m_d^\intercal \mathrm{Q}_d m_d \\
\mathbb{E}\left[w_d(\tilde{\beta}_d - m_d)^\intercal \mathrm{Q}_d (\tilde{\beta}_d - m_d)\right] &= \mathbb{E}\left[w_d \mathbb{E}\left((\tilde{\beta}_d - m_d)^\intercal \mathrm{Q}_d (\tilde{\beta}_d - m_d) \mid w_d\right)\right] \\
&= \mathbb{E}\left[w_d\left(K \cdot w_d + m_d^\intercal \mathrm{Q}_d m_d - w_d m_d^\intercal \mathrm{Q}_d m_d\right)\right] \\
&= \mathbb{E}\left[K \cdot w_d^2 + w_d m_d^\intercal \mathrm{Q}_d m_d - w_d^2 m_d^\intercal \mathrm{Q}_d m_d\right] \\
&= K \cdot \lambda_d + \lambda_d m_d^\intercal \mathrm{Q}_d m_d - \lambda_d m_d^\intercal \mathrm{Q}_d m_d \\
&= K \cdot \lambda_d \\
\mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal\right] &= \int \tilde{\beta}_d \tilde{\beta}_d^\intercal \cdot \lambda_d \cdot \mathcal{N}\left(\tilde{\beta}_d \mid m_d, \mathrm{Q}_d^{-1}\right) d\beta_d = \lambda_d\left(\mathrm{Q}_d^{-1} + m_d m_d^\intercal\right) \\
\mathrm{Var}\left(\tilde{\beta}_d\right) &= \mathbb{E}\left[\tilde{\beta}_d \tilde{\beta}_d^\intercal\right] - \mathbb{E}\left[\tilde{\beta}_d\right]\mathbb{E}\left[\tilde{\beta}_d\right]^\intercal \\
&= \lambda_d\left(\mathrm{Q}_d^{-1} + m_d m_d^\intercal\right) - \lambda_d^2 m_d m_d^\intercal
\end{aligned}
$$