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, mixture of expert models are often used because of their 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 results 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 different alternative 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 $y \mid x$. 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, it 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 using a kernel stick-breaking process (**dunsonpark:08**) or logit stick-breaking prior (**durante:17**) to generate the weights. In the former method, the increased flexibility comes at heavy computational cost, and in the later method, the process from which the weights are generated does not allow for intuitive inference on the covariates. In our proposed model, the covariates enter 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 $\mu_k(x) = x^\intercal \beta_k$ and $\pi_k \propto \exp(x^\intercal \gamma_k)$.

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. We introduce the set of latent variables $\mathbf{Z} = \{z_1, \ldots, z_N\}$, where $z_n \in \mathbb{R}^K$ and $z_{nk} = 1$ 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 simplified form of the marginal likelihood.

$$p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) = \prod_n \prod_k \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right)^{z_{nk}} \tag{3}$$

$$p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) = \prod_n \prod_k \pi_k(x_n)^{z_{nk}} = \prod_n \prod_k \left(\frac{\exp\{x_n^\intercal \gamma_k\}}{\sum_{j=1}^K \exp\{x_n^\intercal \gamma_j\}} \right)^{z_{nk}} \tag{4}$$

Next, we introduce the priors over the parameters $\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}$, where we simplify the calculations by considering conjugate priors.

$$p(\boldsymbol{\gamma}) = \prod_k p(\gamma_k) = \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right) \tag{5}$$

$$p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right) = \prod_k p(\beta_k, \tau_k) = \prod_k p(\beta_k \mid \tau_k) p(\tau_k) = \prod_k \mathcal{N}\left(\beta_k \mid m_0, (\tau_k \Lambda_0)^{-1}\right) \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right) \tag{6}$$

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 is given by

$$\begin{aligned}
\ln\left\{ p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) \right\} &= \ln\left\{ p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \boldsymbol{\gamma}\right) p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right) p\left(\boldsymbol{\gamma}\right) p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right) \right\} \\
&= \sum_n \sum_k z_{nk} \left\{ -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\tau_k - \frac{\tau_k}{2}\left(y_n - x_n^\intercal \beta_k\right)^2 \right\} \\
&+ \sum_n \sum_k z_{nk} \left\{ x_n^\intercal \gamma_k - \ln\left(\sum_{j=1}^K \exp\{x_n^\intercal \gamma_j\} \right) \right\} \\
&+ \sum_k \left\{ -\frac{D}{2}\ln(2\pi) + \frac{D}{2}\ln\tau_k + \ln|\Lambda_0| - \frac{\tau_k}{2}(\beta_k - m_0)^\intercal \Lambda_0 (\beta_k - m_0) \right\} \\
&+ \sum_k \left\{ (a_0 - 1)\ln\tau_k - b_0 \tau_k \right\}
\end{aligned} \tag{7}$$

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

$$q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}\right) = q(\mathbf{Z}) q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}) \tag{8}$$

3.1 Coordinate Ascent Updates

As is standard in variational algorithms, we now seek the sequential updates of the factors in (8). 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 $q(\mathbf{Z})$,

$$\ln q^*(\mathbf{Z}) = E_{-q(\mathbf{Z})}\Big[\ln\left\{p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}, \boldsymbol{\gamma}\right)\right\}\Big] \tag{9}$$

We adopt the convention that the expectation with respect to a negative subscript indicates an expectation taken with repect to the other variational parameters. Ignoring terms that are not functionally dependent on $\mathbf{Z}$, we can exponentiate both sides of (9) to obtain the optimal solution for $q(\mathbf{Z})$,

$$q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{10}$$

For the details in this calculation, refer to the derivations in Appendix A. The form of this discete distribution gives us $\mathbb{E}[z_{nk}] = r_{nk}$. If we consider the quantity $\ln \rho_{nk}$, as defined in (16) in Appendix A, we note that the exact computation of $\ln \rho_{nk}$ involves four expectations taken with respect to the variational distribution $q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma})$. This distribution can be found by considering the expectation of the joint density in (7) taken with respect to $q(\mathbf{Z})$. The resulting variational distribution can be written up to constants as shown below,

$$\begin{aligned}\ln q^*(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}) = &\sum_k \sum_n E_{q(\mathbf{Z})}[z_{nk}] \ln \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right) + \sum_k \ln p\left(\beta_k, \tau_k\right) \\ &+ \sum_k \sum_n E_{q(\mathbf{Z})}[z_{nk}] \left(x_n^\intercal \gamma_k - \ln \sum_{j=1}^{K} \exp\{x_n^\intercal \gamma_j\}\right) + \sum_k \ln \mathcal{N}(\gamma_k \mid 0, \mathrm{I}_D)\end{aligned} \tag{11}$$

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 (25), 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 can obtain the following updates for the remaining variational parameters,

$$q^*(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right) \tag{12}$$

$$q^*(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right) \tag{13}$$

$$q^*(\gamma_k) = \mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right) \tag{14}$$

for $k = 1, \ldots, K$. The derivation for (12) and (13) can be found in Appendix C, and the details for (14) can be found in Appendix B. The definition of each of the variational

parameters can also be found in the corresponding appendices. Now that we have closed form updates 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 (12), (13), and (14). 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.

$$\begin{aligned}\mathcal{L}(q) &= \sum_{z} \int\int\int q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \mathbf{Z}) \ln \left\{ \frac{p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \mathbf{Z})}{q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}, \mathbf{Z})} \right\} d\boldsymbol{\beta} d\boldsymbol{\tau} d\boldsymbol{\gamma} \\ &= \mathbb{E}[\ln p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right)] + \mathbb{E}[\ln p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right)] + \mathbb{E}[\ln p\left(\boldsymbol{\gamma}\right)] + \mathbb{E}[\ln p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)] \\ &\quad - \mathbb{E}[\ln q(\mathbf{Z})] - \mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})] - \mathbb{E}[\ln q(\boldsymbol{\gamma})]\end{aligned} \tag{15}$$

Each of the seven expectations can be expressed as follows. Details for each calculation can be found in Appendix D.

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

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

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

$$\begin{aligned}\mathbb{E}[\ln p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)] &= \left(a_0 + \frac{D}{2} - 1\right) \sum_{k} \psi(a_k) - \psi(b_k) - \frac{1}{2} \sum_{k} \frac{a_k}{b_k} (m_k - m_0)^\intercal \Lambda_0 (m_k - m_0) + \operatorname{tr}\left(\Lambda_0 \mathrm{V}_k^{-1}\right) \\ &\quad - K \left(\frac{D}{2} \ln(2\pi) - \ln|\Gamma_0| - a_0 \ln b_0 + \ln \Gamma(\alpha_0) \right)\end{aligned}$$

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

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

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

4. Algorithm

Using the updates discussed in the previous section, we can formalize the variational algorithm below.

Algorithm 1 CAVI for Conditional Density Estimation

Result: Write here the result

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 $q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}\right) = q(\mathbf{Z})q(\boldsymbol{\beta}, \boldsymbol{\tau}, \boldsymbol{\gamma}) = q(\mathbf{Z}) \prod_k q(\beta_k, \tau_k)q(\gamma_k)$;

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

  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 $\eta_k \leftarrow \sum_n r_{nk}\big[\frac{1}{2} + 2\lambda(\xi_{nk})\alpha_n\big] x_n$

    Set $\mu_k \leftarrow \mathrm{Q}_k^{-1}\eta_k$

    Set $\mathrm{V}_k \leftarrow \sum_n r_{nk} x_n x_n^\intercal + \Lambda_0$

    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 $b_k \leftarrow b_0 + \frac{1}{2}[\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - \zeta_k^\intercal \mathrm{V}_k^{-1}\zeta_k]$

  end

  Compute ELBO using updated parameters

end

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

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 $\mathbf{Z}$,

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

where we have defined

$$\begin{aligned}\ln \rho_{nk} = &-\frac{1}{2}\ln(2\pi) + \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\tau})}[\ln \tau_k] - \frac{1}{2}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k (y_n - x_n^\intercal \beta_k)^2] \\ &+ x_n^\intercal \mathbb{E}_{q(\boldsymbol{\gamma})}[\gamma_k] - \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln \left(\sum_j \exp\{x_n^\intercal \gamma_j\} \right) \right]\end{aligned} \tag{16}$$

Exponentiating and normalizing, we have

$$q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{17}$$

For the discrete distribution $q^*(\mathbf{Z})$ given in (17) 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 (16).

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 $q^*(\boldsymbol{\tau})$).

$$E_{q(\boldsymbol{\tau})}[\ln \tau_k] = \psi(a_k) - \psi(b_k) \tag{18}$$

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

$$\begin{aligned}\mathbb{E}_{q(\boldsymbol{\beta},\boldsymbol{\tau})}\left[\tau_k (y_n - x_n^\intercal \beta_k)^2\right] &= \mathbb{E}\left[\tau_k \left(y_n - m_k^\intercal x_n x_n^\intercal m_k + \mathrm{tr}\left(x_n x_n^\intercal (\tau_k V_k)^{-1} \right) - 2 y_n x_n^\intercal m_k \right) \right] \\ &= \frac{a_k}{b_k}\left(y_n^2 + m_k^\intercal x_n x_n^\intercal m_k \right) + \mathrm{tr}\left(x_n x_n^\intercal V_k^{-1} \right) \\ &= \frac{a_k}{b_k}(y_n + m_k^\intercal x_n)^2 + x_n^\intercal V_k^{-1} x_n\end{aligned} \tag{19}$$

From the expression derived in (24) of Appendix B, we have $q^*(\gamma_k) = \mathcal{N}(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1})$, then we have

$$\mathbb{E}_{q(\gamma_k)}[\gamma_k] = \mu_k \tag{20}$$

Using the bound discussed in Appendix B, equation (23), we can then compute the following expectation with respect to $q^*(\boldsymbol{\gamma})$.

$$\begin{aligned}
&\mathbb{E}_{q(\boldsymbol{\gamma})}\left[\ln\left(\sum_j^K \exp\{x_n^\intercal \gamma_j\}\right)\right] \\
&\approx \mathbb{E}_{q(\boldsymbol{\gamma})}\left[\alpha_n + \sum_{j=1}^K \frac{x_n^\intercal\gamma_j - \alpha_n + \xi_{nj}}{2} + \lambda(\xi_{nj})\left((x_n^\intercal\gamma_j - \alpha_n)^2 - \xi_{nj}^2\right) + \log\left(1+e^{\xi_{nj}}\right)\right] \\
&= \alpha_n + \sum_j^K \frac{1}{2}\left(x_n^\intercal\mu_j - \alpha_n + \xi_{nj}\right) + \lambda(\xi_{nj})\left((x_n^\intercal\mu_j - \alpha_k)^2 - \xi_{nj}^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n\right) + \log(1+e^{\xi_{nj}})
\end{aligned} \tag{21}$$

Gathering the results in (18), (19), (20), and (21), and substituting these into (16), 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:

$$\log\left(\sum_{j=1}^K \exp\{x_n^\intercal\gamma_j\}\right) \leq \alpha_n + \sum_{j=1}^K \log\left(1+\exp\{x_n^\intercal\gamma_j - \alpha_n\}\right) \tag{22}$$

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:

$$\log\left(\sum_{j=1}^K \exp\{x_n^\intercal\gamma_j\}\right) \leq \alpha_n + \sum_{j=1}^K \frac{x_n^\intercal\gamma_j - \alpha_n + \xi_{nj}}{2} + \lambda(\xi_{nj})\left((x_n^\intercal\gamma_j - \alpha_n)^2 - \xi_{nj}^2\right) + \log\left(1+e^{\xi_{nj}}\right) \tag{23}$$

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 (22), thus allowing us to obtain a closed form for the

variational distribution. Note that all of the equalities above are written up to constants.

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

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

$$\begin{aligned}
\mu_k &= \mathrm{Q}_k^{-1}\eta_k \\
\eta_k &= \sum_n r_{nk}\left(\frac{1}{2} + 2\lambda(\xi_{nj})\alpha_n\right)x_n \\
\mathrm{Q}_k &= \mathrm{I}_D + 2\sum_n r_{nk}\lambda(\xi_{nk})x_n x_n^\intercal
\end{aligned} \tag{24}$$

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),

$$\begin{aligned}
\ln q^*(\beta_k, \tau_k) = &\sum_n -\frac{1}{2}r_{nk}\tau_k\left(y_n^2 + \beta_k^\intercal x_n x_n^\intercal\beta_k - 2y_n\beta_k^\intercal x_n\right) + \frac{r_{nk}}{2}\ln\tau_k + \frac{D}{2}\ln\tau_k \\
&- \frac{\tau_k}{2}\left(\beta_k^\intercal\Lambda_0\beta_k + m_0^\intercal\Lambda_0 m_0 - 2\beta_k^\intercal\Lambda_0 m_0\right) + (a_0 - 1)\ln\tau_k - b_0\tau_k
\end{aligned} \tag{25}$$

We first consider terms on the right hand side of (25) that depend on β_k to find $\ln q^\star(\beta_k \,|\, \tau_k)$, giving

$$\ln q^\star(\beta_k \mid \tau_k) = -\frac{\tau_k}{2}\beta_k^\intercal\Big[\sum_n r_{nk}x_n x_n^\intercal + \Lambda_0\Big]\beta_k + \tau_k\beta_k^\intercal\Big[\sum_n r_{nk}y_n x_n + \Lambda_0 m_0\Big] \tag{26}$$

$$q^*(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right) \tag{27}$$

$$\begin{aligned} \mathrm{V}_k &= \sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 \\ \zeta_k &= \sum_n r_{nk} y_n x_n + \Lambda_0 m_0 \\ m_k &= \mathrm{V}_k^{-1} \zeta_k \end{aligned} \tag{28}$$

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 (25) and (27). Note that equality below is written up to constants, keeping only terms involving τ_k.

$$\begin{aligned} \ln q^*(\tau_k) = (a_0 + N_k - 1)\ln \tau_k - \tau_k \Bigg\{ & b_0 + \frac{1}{2}\left(\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - m_k^\intercal \mathrm{V}_k m_k\right) \\ & + \frac{1}{2}\beta_k^\intercal \left(\sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 - \mathrm{V}_k\right)\beta_k \\ & - 2\beta_k^\intercal \left(\sum_n r_{nk} y_n x_n + \Lambda_0 m_0 - \mathrm{V}_k m_k\right) \Bigg\} \end{aligned} \tag{29}$$

Exponentiating, we arrive at the following distribution

$$q^*(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right) \tag{30}$$

where we have defined

$$\begin{aligned} a_k &= a_0 + N_k \\ b_k &= b_0 + \frac{1}{2}\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - \zeta_k^\intercal \mathrm{V}_k^{-1} \zeta_k \end{aligned} \tag{31}$$

The expression for b_k arises by noting that the three following simplifications for the summation terms in the coefficient of τ_k in (29),

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

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 (28), and the third equality holds by observing from (28) that $\mathrm{V}_k m_k = \zeta_k$

Appendix D. Variational Lower Bound

As seen in (15) 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.

$$\begin{aligned}
\mathbb{E}[\ln p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right)] &= \sum_n \sum_k \mathbb{E}[z_{nk}] \mathbb{E}\Big[\ln \mathcal{N}\left(y_n \mid x_n^\intercal \beta_k, \tau_k^{-1}\right) \Big] \\
&= -\frac{1}{2} \sum_n \sum_k r_{nk} \Big\{ \ln(2\pi) - \mathbb{E}[\ln \tau_k] + \mathbb{E}[\tau_k \left(y_n - x_n^\intercal \beta_k\right)^2] \Big\} \\
&= -\frac{1}{2} \sum_n \sum_k r_{nk} \Big\{ \ln(2\pi) - (\psi(a_k) + \psi(b_k)) + x_n^\intercal \mathrm{V}_k^{-1} x_n \\
&\qquad + \frac{a_k}{b_k} (y_n - x_n^\intercal m_k)^2 \Big\}
\end{aligned} \tag{32}$$

$$\begin{aligned}
\mathbb{E}[\ln p\left(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\gamma}\right)] &= \mathbb{E}[z_{nk}] \left(\mathbb{E}[x_n^\intercal \gamma_k] - \mathbb{E}\Big[\ln \sum_j \exp\{x_n^\intercal \gamma_j\} \Big] \right) \\
&\approx \sum_n \sum_k r_{nk} \left(x_n^\intercal \mu_k - \alpha_n - \varphi_n \right)
\end{aligned} \tag{33}$$

where $\varphi_n = \sum_{j=1}^{K} \frac{1}{2} \left(x_n^\intercal \mu_j - \alpha_n + \xi_{nj} \right) + \lambda(\xi_{nj}) \left((x_n^\intercal \mu_j - \alpha_k)^2 - \xi_{nj}^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n \right) + \log(1 + e^{\xi_{nj}})$. Here, we make use of the result in (21), where we take the expectation of the upper bound previously derived.

$$\begin{aligned}
\mathbb{E}[\ln p\left(\boldsymbol{\gamma}\right)] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right) \Big] \\
&= -\frac{K \cdot D}{2} \ln(2\pi) - \frac{1}{2} \sum_k \mu_k^\intercal \mu_k
\end{aligned} \tag{34}$$

$$\begin{aligned}
\mathbb{E}[\ln p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\beta_k \mid m_0, (\tau_k \Lambda_0)^{-1}\right) \Big] + \mathbb{E}\Big[\ln \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right) \Big] \\
&= \left(a_0 + \frac{D}{2} - 1 \right) \sum_k \psi(a_k) - \psi(b_k) \\
&\quad - \frac{1}{2} \sum_k \frac{a_k}{b_k} (m_k - m_0)^\intercal \Lambda_0 (m_k - m_0) + \mathrm{tr}\left(\Lambda_0 \mathrm{V}_k^{-1}\right) \\
&\quad - K \left(\frac{D}{2} \ln(2\pi) - \ln|\Gamma_0| - a_0 \ln b_0 + \ln \Gamma(\alpha_0) \right)
\end{aligned} \tag{35}$$

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 $E[z_{nk}] = r_{nk}$.

$$\mathbb{E}[\ln q(\mathbf{Z})] = \sum_n \sum_k r_{nk} \ln r_{nk} \tag{36}$$

$$\begin{aligned}\mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})] &= \sum_k \mathbb{E}\Big[\ln \mathcal{N}\left(\beta_k \mid m_k, (\tau_k V_k)^{-1}\right)\Big] + \mathbb{E}\Big[\ln \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right)\Big] \\ &= \sum_k \left(\frac{D}{2} + a_k - 1\right)\left[\psi(a_k) - \psi(b_k)\right] + a_k \ln b_k - a_k - \ln \Gamma(a_k) \\ &\quad - K\left(\frac{D}{2}\ln(2\pi) - \ln|\mathrm{V}_k| + \frac{D}{2}\right)\end{aligned} \tag{37}$$

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.

$$\begin{aligned}\mathbb{E}[\ln q(\boldsymbol{\gamma})] &= \sum_k \mathbb{E}\Big[\mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right)\Big] \\ &= -\frac{KD}{2}(\ln(2\pi) + 1) + \sum_k \ln|\mathrm{Q}_k|\end{aligned} \tag{38}$$

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

sample