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 $y \,|\, x$. A common appraoch for doing this is to model the density using a mixture of gaussians, such as the following,

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

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 $\mathbf{y} = \{y_1, \dots, y_N\}, \mathbf{X} = \{x_1, \dots, x_N\}$, where each $x_n \in \mathbb{R}^D$. Let $\boldsymbol{\beta} = \{\beta_1, \dots, \beta_K\}$ and $\boldsymbol{\gamma} = \{\gamma_1, \dots, \gamma_k\}$ denote the D-dimensional coefficient vectors in the mixture weights and in gaussian mixture components, respectively. Finally, let $\boldsymbol{\tau} = \{\tau_1, \dots, \tau_k\}$ denote the precision parameters. We introduce the set of latent variables $\mathbf{Z} = \{z_1, \dots, z_N\}$, 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 $\mathbf{Z}$, we have the following simplified form of the marginal likelihood.

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

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

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 distibution 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). Taking the expectation with respect to the other variational parameters, we can derive 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 $q(\mathbf{Z})$,

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

See Appendix A for the details in this calculation.
It remains to consider the factor $q(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma})$. Taking the expectation with respect to $q(\mathbf{Z})$, we have the following equality written up to constants,

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

From the expression above, we see that the optimal distribution has a sum involving only the γ_k's and a sum involving only the (β_k,τ_k)'s, which implies $q(\boldsymbol{\beta},\boldsymbol{\tau},\boldsymbol{\gamma}) = \left[\prod_k q(\beta_k,\tau_k)\right]\prod_k q(\gamma_k)$. With this factorization in mind, we can obtain the following updates for the remaining variational parameters,

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

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

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

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

3.2 Evidence Lower Bound (ELBO)

4. Algorithm

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

Algorithm 1 Variational Approximation for Gaussian Mixture

```
procedure MyProcedure
    $stringlen \leftarrow$ length of $string$
    $i \leftarrow patlen$
    $top$:
    if $i > stringlen$ then return false
    $j \leftarrow patlen$
    $loop$:
    if $string(i) = path(j)$ then
        $j \leftarrow j - 1$.
        $i \leftarrow i - 1$.
        goto $loop$.
        close;
    $i \leftarrow i + \max(delta_1(string(i)), delta_2(j))$.
    goto $top$.
```

Appendix A.

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

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

where we have defined

$$\begin{aligned} \ln \rho_{nk} = & -\frac{1}{2}\ln(2\pi) + \frac{1}{2}E_{q(\boldsymbol{\tau})}[\ln \tau_k] - \frac{1}{2}E_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^\intercal \beta_k)^2] \\ & + x_n^\intercal E_{q(\boldsymbol{\gamma})}[\gamma_k] - E_{q(\boldsymbol{\gamma})}\left[\ln\left(\sum_j \exp\{x_n^\intercal \gamma_j\}\right)\right] \end{aligned} \tag{15}$$

Exponentiating and normalizing, we have

$$q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{16}$$

For the discrete distribution $q^*(\mathbf{Z})$ given in (16) 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 (15).

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

$$E_{q(\boldsymbol{\tau})}[\ln \tau_k] = \psi(a_k) - \psi(b_k) \tag{17}$$

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

$$\begin{aligned} E_{q(\boldsymbol{\beta},\boldsymbol{\tau})}\left[\tau_k(y_n - x_n^\intercal \beta_k)^2\right] &= E\left[\tau_k \left(y_n - m_k^\intercal x_n x_n^\intercal m_k + \mathrm{tr}\left(x_n x_n^\intercal \left(\tau_k V_k\right)^{-1}\right) - 2 y_n x_n^\intercal m_k\right)\right] \\ &= \frac{a_k}{b_k}\left(y_n^2 + m_k^\intercal x_n x_n^\intercal m_k\right) + \mathrm{tr}\left(x_n x_n^\intercal V_k^{-1}\right) \\ &= \frac{a_k}{b_k}(y_n + m_k^\intercal x_n)^2 + x_n^\intercal V_k^{-1} x_n \end{aligned} \tag{18}$$

From the expression derived in (23) of Appendix B, we have $q^*(\gamma_k) = \mathcal{N}(\gamma_k \,|\, \mu_k, \mathrm{Q}_k^{-1})$, then we have

$$E_{q(\gamma_k)}[\gamma_k] = \mu_k \tag{19}$$

Using the bound discussed in Appendix B, equation (22), we can then compute the following expectation with respect to $q^*(\boldsymbol{\gamma})$.

$$\begin{aligned} &\mathrm{E}_{q(\boldsymbol{\gamma})}\left[\ln\left(\sum_j^K \exp\{x_n^\intercal \gamma_j\}\right)\right] \\ &\approx \mathrm{E}_{q(\boldsymbol{\gamma})}\left[\alpha_n + \sum_{j=1}^K \frac{x_n^\intercal \gamma_j - \alpha_n + \xi_{nj}}{2} + \lambda(\xi_{nj})\left((x_n^\intercal \gamma_j - \alpha_n)^2 - \xi_{nj}^2\right) + \log\left(1 + e^{\xi_{nj}}\right)\right] \\ &= \alpha_n + \sum_j^K \frac{1}{2}\left(x_n^\intercal \mu_j - \alpha_n + \xi_{nj}\right) + \lambda(\xi_{nj})\left((x_n^\intercal \mu_j - \alpha_k)^2 - \xi_{nj}^2 + x_j^\intercal \mathrm{Q}_k^{-1} x_j\right) + \log(1 + e^{\xi_{nj}}) \end{aligned} \tag{20}$$

Gathering the results in (17), (18), (19), and (20), and substituting these into (15), we can compute $E[z_{nk}] = r_{nk}$ in closed form.

Appendix B.

For the variational distribution for $\gamma_k, k = 1, \ldots, K$, we first note the following bound given by **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:

$$\log\left(\sum_{j=1}^K \exp\{x_n^\intercal \gamma_j\}\right) \leq \alpha_n + \sum_{j=1}^K \log\left(1 + \exp\{x_n^\intercal \gamma_j - \alpha_n\}\right) \tag{21}$$

If we then use the bound from **jj:2001**

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

then we arrive at the following bound:

$$\log\left(\sum_{j=1}^{K}\exp\{x_n^\intercal\gamma_j\}\right) \leq \alpha_n+\sum_{j=1}^{K}\frac{x_n^\intercal\gamma_j-\alpha_n+\xi_{nj}}{2}+\lambda(\xi_{nj})\left((x_n^\intercal\gamma_j-\alpha_n)^2-\xi_{nj}^2\right)+\log\left(1+e^{\xi_{nj}}\right) \tag{22}$$

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 (21), thus allowing us to obtain a closed form for the variational distribution. Note that all of the equalities above are written up to constants.

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

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

$$\begin{aligned}
\mu_k &= \mathrm{Q}_k^{-1}\eta_k\\
\eta_k &= \sum_n r_{nk}\left(\frac{1}{2}+2\lambda(\xi_{nj})\alpha_n\right)x_n\\
\mathrm{Q}_k &= \mathrm{I}_D+2\sum_n r_{nk}\lambda(\xi_{nk})x_n x_n^\intercal
\end{aligned} \tag{23}$$

The additional parameters introduced in the two upper bounds can be updated using the following equations

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

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

Appendix C.

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

$$\begin{aligned}\ln q^{*}(\beta_k, \tau_k) = &\sum_n -\frac{1}{2} r_{nk} \tau_k \left(y_n^2 + \beta_k^\intercal x_n x_n^\intercal \beta_k - 2 y_n \beta_k^\intercal x_n \right) + \frac{r_{nk}}{2} \ln \tau_k + \frac{D}{2} \ln \tau_k \\ &- \frac{\tau_k}{2} \left(\beta_k^\intercal \Lambda_0 \beta_k + m_0^\intercal \Lambda_0 m_0 - 2 \beta_k^\intercal \Lambda_0 m_0 \right) + (a_0 - 1) \ln \tau_k - b_0 \tau_k\end{aligned} \tag{24}$$

We first consider terms on the right hand side of (24) that depend on $beta_k$ to find $\ln q^{\star}(\beta_k \mid \tau_k)$, giving

$$\ln q^{\star}(\beta_k \mid \tau_k) = -\frac{\tau_k}{2} \beta_k^\intercal \Big[\sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 \Big] \beta_k + \tau_k \beta_k^\intercal \Big[\sum_n r_{nk} y_n x_n + \Lambda_0 m_0 \Big] \tag{25}$$

$$q^{*}(\beta_k \mid \tau_k) = \mathcal{N}\left(m_k, (\tau_k \mathrm{V}_k)^{-1} \right) \tag{26}$$

$$\begin{aligned} m_k &= \mathrm{V}_k^{-1} b_k \\ \mathrm{V}_k &= \sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 \\ b_k &= \sum_n r_{nk} y_n x_n + \Lambda_0 m_0 \end{aligned} \tag{27}$$

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 (24) and (26). Note that equality below is written up to constants, keeping only terms involving τ_k.

$$\begin{aligned}\ln q^{*}(\tau_k) = (a_0 + N_k - 1) \ln \tau_k - \tau_k \Bigg\{ &b_0 + \frac{1}{2} \left(\sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - m_k^\intercal \mathrm{V}_k m_k \right) \\ &+ \frac{1}{2} \beta_k^\intercal \left(\sum_n r_{nk} x_n x_n^\intercal + \Lambda_0 - \mathrm{V}_k \right) \beta_k \\ &- 2 \beta_k^\intercal \left(\sum_n r_{nk} y_n x_n + \Lambda_0 m_0 - \mathrm{V}_k m_k \right) \Bigg\}\end{aligned} \tag{28}$$

Exponentiating, we arrive at the following distribution

$$q^{*}(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k \right) \tag{29}$$

where we have defined

$$\begin{aligned} a_k &= a_0 + N_k \\ b_k &= b_0 + \frac{1}{2} \sum_n r_{nk} y_n^2 + m_0^\intercal \Lambda_0 m_0 - b_k^\intercal \mathrm{V}_k^{-1} b_k \end{aligned} \tag{30}$$

The expression for b_k arises by noting that the three following simplifications for the summation terms in the coefficient of τ_k in (28),

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

where the first equality holds by expanding $m_k^\intercal \mathrm{V}_k m_k = b_k^\intercal \left(\mathrm{V}_k^{-1}\right)^\intercal \mathrm{V}_k \mathrm{V}_k^{-1} b_k = b_k^\intercal \mathrm{V}_k^{-1} b_k$. The second quality holds by recalling the definition of V_k in (27), and the third equality holds by observing from (27) that $\mathrm{V}_k m_k = b_k$

Appendix D.

sample