Variational Approach for Bayesian Density Regression

Eric Chuu

Texas A&M University

ericchuu@tamu.com

December 11, 2018

Overview

1. Introduction and Setup
2. Main Problem
   - Computational Bottlenecks
3. Variational Approach
   - CAVI Updates
   - Bouchard Bound for Problematic Quantity
   - Bouchard Bound for Problematic Quantity

Problem Setup

- Given data $(y_n, x_n)_{n=1}^{N}$ and we want to estimate the density of $y \mid x$
- Gaussian mixture models are a common choice due to flexibility

$$f(y \mid x) = \sum_{k} \pi_k \mathcal{N}(y \mid \mu_k(x), \tau_k^{-1})$$

- Covariate-independent weights have limited flexibility in practice

Covariate Dependent Weights

- Allow for covariate dependence in the mixing weights

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

- Existing methods that use kernel stick breaking process (Dunson and Park, 2008), logit stick breaking prior (Rigon and Durante, 2017)
- We allow covariates to enter the weights via a logit link function

$$\pi_k(x) \propto \exp\{x^\intercal \gamma_k\}$$

Notation

- Observed Data: $\mathbf{y} = \{y_1, \ldots, y_N\}, \mathbf{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^D$
- Guassian Componenets: $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_K\}, \boldsymbol{\tau} = \{\tau_1, \ldots, \tau_k\}$

$$f(y \mid x) = \sum_k \pi_k(x)\, \mathcal{N}(y \mid x^\intercal \beta_k, \tau_k^{-1})$$

- coefficient vectors in the weights: $\boldsymbol{\gamma} = \{\gamma_1, \ldots, \gamma_k\}$
- Latent variables: $\mathbf{Z} = \{z_1, \ldots, z_N\}$
  - $z_n \in \mathbb{R}^K$ and $z_{nk} = 1$ iff y_n belongs to the k-th cluster

Prior Specification

With the introduction of latent variables, we have the simplified conditional distribution,

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

The conditional distribution of $\mathbf{Z}$ given $\mathbf{X}, \gamma$

$$p\left(\mathbf{Z} \mid \mathbf{X}, \gamma\right) = \prod_{n} \prod_{k} \pi_k(x_n)^{z_{nk}} = \prod_{n} \prod_{k} \left(\frac{\exp\{x_n^{\intercal} \gamma_k\}}{\sum_{j=1}^{K} \exp\{x_n^{\intercal} \gamma_j\}} \right)^{z_{nk}} \tag{2}$$

Priors over $\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma$

- $p(\gamma) = \prod_k p(\gamma_k) = \prod_k \mathcal{N}\left(\gamma_k \mid 0, \mathrm{I}_D\right)$
- $p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right) = p\left(\boldsymbol{\beta} \mid \boldsymbol{\tau}\right) p(\boldsymbol{\tau}) = \prod_k \mathcal{N}\left(\beta_k \mid m_0, (\tau_k \Lambda_0)^{-1}\right) \mathrm{Ga}\left(\tau_k \mid a_0, b_0\right)$

General Approach

Full joint distribution of all random variables

$$p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) = p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) p\left(\mathbf{Z} \mid \mathbf{X}, \gamma\right) p\left(\gamma\right) p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)$$

Formulate an approximating distribution

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

Derive update equations for the variational distributions

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

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

Variational Approximation

Joint Distribution of all random variables

$$\ln p\left(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right) = \sum_n \sum_k z_{nk} \left\{ -\frac{1}{2}\ln(2\pi) + \frac{1}{2}\ln\tau_k - \frac{\tau_k}{2}\left(y_n - x_n^{\mathsf{T}}\beta_k\right)^2 \right\}$$

$$+ \sum_n \sum_k z_{nk} \left\{ x_n^{\mathsf{T}}\gamma_k - \ln\left(\sum_{j=1}^{K} \exp\{x_n^{\mathsf{T}}\gamma_j\}\right) \right\}$$

$$+ \sum_k \left\{ -\frac{D}{2}\ln(2\pi) + \frac{D}{2}\ln\tau_k + \ln|\Lambda_0| - \frac{\tau_k}{2}(\beta_k - m_0)^{\mathsf{T}}\Lambda_0(\beta_k - m_0) \right\}$$

$$+ \sum_k \left\{ (a_0 - 1)\ln\tau_k - b_0\tau_k \right\}$$

Approximating Distribution

$$q\left(\mathbf{Z}, \boldsymbol{\beta}, \boldsymbol{\tau}, \gamma\right) = q(\mathbf{Z})q(\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma) \tag{3}$$

Example calculation for $q(\mathbf{Z})$

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

Exponentiating and normalizing, we arrive at the optimal variational distribution. Then, $\mathbb{E}[z_{nk}] = r_{nk}$.

$$
q^*(\mathbf{Z}) = \prod_n \prod_k r_{nk}^{z_{nk}}, \quad r_{nk} = \frac{\rho_{nk}}{\sum_j \rho_{nj}} \tag{4}
$$

Example calculation for $q(\mathbf{Z})$

Consider the quantity $\ln \rho_{nk}$:

- $E_{q(\boldsymbol{\tau})}[\ln \tau_k]$
- $E_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^{\intercal}\beta_k)^2]$
- $E_{q(\gamma)}[\gamma_k]$
- $E_{q(\gamma)}\left[\ln\left(\sum_j \exp\{x_n^{\intercal}\gamma_j\}\right)\right]$

Variational Updates

Our choice of conjugate families tells us that we will have the following form of the variational distributions, for $k = 1, \ldots, K$,

$$q^*(\gamma_k) = \mathcal{N}\left(\gamma_k \mid \mu_k, \mathrm{Q}_k^{-1}\right) \tag{5}$$

$$q^*(\beta_k \mid \tau_k) = \mathcal{N}\left(\beta_k \mid m_k, (\tau_k \mathrm{V}_k)^{-1}\right) \tag{6}$$

$$q^*(\tau_k) = \mathrm{Ga}\left(\tau_k \mid a_k, b_k\right) \tag{7}$$

Example calculation for $q(\mathbf{Z})$

Consider the quantity $\ln \rho_{nk}$:

- $E_{q(\boldsymbol{\tau})}[\ln \tau_k] = \psi(a_k) - \psi(b_k)$
- $E_{q(\boldsymbol{\beta},\boldsymbol{\tau})}[\tau_k(y_n - x_n^\intercal \beta_k)^2] = \frac{a_k}{b_k}(y_n + m_k^\intercal x_n)^2 + x_n^\intercal \mathrm{V}_k^{-1} x_n$
- $E_{q(\gamma)}[\gamma_k] = \mu_k$
- $E_{q(\gamma)}\left[\ln\left(\sum_j \exp\{x_n^\intercal \gamma_j\}\right)\right] = \;\ldots \text{ oh no!}$

Bouchard's Bound for Problematic Quantity

Upper bound on the sum of exponentials

$$\sum_{j=1}^{K} e^{t_j} \leq \prod_{j=1}^{K} (1 + e^{t_j})$$

Tangential Bound (Jaakkola and Jordan, 1996)

For $x \in \mathbb{R}, \alpha \in \mathbb{R}, \xi \in [0, \infty)$

$$\log(1 + e^x) \leq \lambda(\xi)(x^2 - \xi^2) + \frac{x - \xi}{2} + \log(1 + e^{\xi})$$

Taking log the first bound and setting $t_j = x_n^\intercal \gamma_j - \alpha_n$, we have

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

$$+ \log\left(1 + e^{\xi_{nj}}\right)$$

Bouchard's Bound for Problematic Quantity

Upper Bound for the problematic expectation

$\mathrm{E}_{q(\gamma)}\left[\ln\left(\sum_j^K \exp\{x_n^\intercal \gamma_j\}\right)\right]$

$$\leq \alpha_n + \sum_j^K \lambda(\xi_{nj})\left((x_n^\intercal \mu_j - \alpha_k)^2 - \xi_{nj}^2 + x_n^\intercal \mathrm{Q}_k^{-1} x_n\right) + \log(1 + e^{\xi_{nj}})$$
$$+ \frac{1}{2}\left(x_n^\intercal \mu_j - \alpha_n + \xi_{nj}\right)$$

Two new variational parameters that need to be updated

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

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

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

Variational Algorithm

Input

1. Data $(y_n, x_n)_{n=1}^N$
2. Number of components, K
3. Prior mean, precision for coefficients vectors, $\boldsymbol{\beta}_{1:K}$
4. Prior shape, rate parameters for precision parameters, $\tau_{1:K}$

Output

1. A variational density,

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

2. Fully specified by the variational parameters

Algorithm. CAVI for Conditional Density Estimation

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

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

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

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

    end

  end

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

    for $k \in \{1, \dots, K\}$ **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 $\alpha_n \leftarrow \Big[\frac{1}{2}\left(\frac{K}{2} - 1\right) + \sum_k \lambda(\xi_{nk}) \mu_k^\intercal x_n\Big] \Big/ \Big[\sum_k \lambda(\xi_{nk})\Big]$

  end

  `/** Remaining Variational Updates on Next Slide **/`

end

Algorithm. CAVI for Conditional Density Estimation

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

$\vdots$

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

$\mathrm{Q}_k \leftarrow \mathrm{I}_D + 2\sum_n r_{nk}\lambda(\xi_{nk}) x_n x_n^\intercal$ /* gamma_k cov */

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

$\mu_k \leftarrow Q_k^{-1}\eta_k$ /* gamma_k mean */

$\mathrm{V}_k \leftarrow \sum_n r_{nk} x_n x_n^\intercal + \Lambda_0$ /* beta_k cov */

$\zeta_k \leftarrow \sum_n r_{nk} y_n x_n + \Lambda_0 m_0$

$m_k \leftarrow \mathrm{V}_k^{-1}\zeta_k$ /* beta_k mean */

$a_k \leftarrow a_0 + N_k$ /* tau_k shape */

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

end

Compute ELBO using updated parameters

end

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

Monitoring Convergence

At the end of each iteration, we compute the ELBO:

$$\mathcal{L}(q) = \sum_{z} \int \int \int q(\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z}) \ln \left\{ \frac{p(\mathbf{y}, \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z})}{q(\boldsymbol{\beta}, \boldsymbol{\tau}, \gamma, \mathbf{Z})} \right\} d\boldsymbol{\beta} d\boldsymbol{\tau} d\gamma$$

7 quantities to calculate:

- $\mathbb{E}[\ln p\left(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \boldsymbol{\tau}, \mathbf{Z}\right)]$
- $\mathbb{E}[\ln p\left(\mathbf{Z} \mid \mathbf{X}, \gamma\right)]$
- $\mathbb{E}[\ln p\left(\gamma\right)]$
- $\mathbb{E}[\ln p\left(\boldsymbol{\beta}, \boldsymbol{\tau}\right)]$
- $\mathbb{E}[\ln q(\mathbf{Z})]$
- $\mathbb{E}[\ln q(\boldsymbol{\beta}, \boldsymbol{\tau})]$
- $\mathbb{E}[\ln q(\gamma)]$

References

Please see report submission for updated set of references.

Thank you!