An Exploration of Variational Inference

Evan Glas[1]
[1]Duke University

Abstract

Bayesian inference is a systematic framework to update preexisting beliefs given new information. In terms of Bayes rule, this framework, enables us to compute a *posterior distribution* over some unknown quantity given a *prior distribution* and a *likelihood function* parametrized by that quantity. A central goal in Bayesian inference is the efficient approximation of potentially intractable posterior distributions. Variational inference (VI) is a method that seeks to achieve this goal by posing posterior estimation as an optimization problem. In this project, I demonstrate the utility of VI by fitting a Gaussian Mixture Model to synthetic data. I first describe the mathematical underpinnings of VI as follows from Bayes rule. I follow with a brief description of applications of variational inference. I then provide a mathematical derivation of a well-known optimization algorithm, Coordinate Ascent Variational Inference (CAVI). Finally, I implement this algorithm in code and visualize the model-fitting process.

Background: Bayes Rule

At the core of Bayesian inference lies Bayes rule, a mathematical expression describing how to update a set of beliefs given new information:

$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)} \tag{1}$$

Equation (1) comprises four components: the posterior $P(\theta|X)$, the likelihood $P(X|\theta)$, the prior $P(\theta)$, and the evidence $P(X)$. In this expression, θ represents some quantity of interest while X represents the new information. Often, the likelihood function and prior are known quantities in this equation. One may specify the likelihood according to a generative model of the data given θ, and the prior may be chosen to impose varying degrees of confidence on the true value of θ. However, the evidence, $P(X)$ may be difficult to compute.

$$P(X) = \int_\theta P(X|\theta)P(\theta)d\theta \tag{2}$$

We may express the evidence as a marginalization over the joint $P(X, \theta)$. However, Equation (2) may be intractable under many likelihood prior pairs. It may be difficult to even approximate the evidence, especially when working in high dimensions. Often, one may choose to restrict their model such that the likelihood and prior form a conjugate pair and yield a known posterior. If such an approach is unsuitable, there also exist sampling methods to approximate the posterior via repeated simulation. However, sampling-based approaches may require significant computational resources before convergence.

Background: Evidence Lower Bound (ELBO)

VI poses posterior estimation as an optimization problem in which one seeks to maximize the closeness between a proposal distribution, Q, and the true posterior (target) distribution, P. The term "variational" stems from *calculus of variations*, the study of the optimization of functionals, or "function-valued of functions". This phrase applies as VI seeks find an optimal distribution (i.e., a function over a given unknown quantity) rather than an optimal quantity itself. The objective function of VI is stated as follows:

$$\min_Q D_{KL}(Q||P) = \min_Q \int_\theta Q(\theta) \log \frac{Q(\theta)}{P(\theta|X)} d\theta \tag{3}$$

In Equation (3), D_{KL} is the Kullback-Leibler (KL) divergence, a measure of distance between two probability distributions. This minimization problem has an equivalent formulation as a maximization problem. Using the property that $D_{KL}(Q||P) \geq 0$ (over P and Q where the KL divergence is defined) we have the following:

$$\log P(X) \geq \int_\theta Q(\theta) \log \frac{P(X, \theta)}{Q(\theta)} d\theta = \mathbb{E}_Q \log \frac{P(X, \theta)}{Q(\theta)} \tag{4}$$

The right-hand side of Equation (5) is known as the evidence lower bound (ELBO) as it provides a lower bound on the log of the evidence, $P(X)$. Maximizing the ELBO over Q is equivalent to minimizing the KL divergence between $Q(\theta)$ and $P(\theta|X)$. VI techniques generally involve some optimization procedure to maximize the ELBO. One may use the ELBO to evaluate the progress of the model fitting process, as well as a tool to compare different models after being fit. A higher ELBO indicates a proposal distribution that better matches the prior distribution and observed data.

Coordinate Ascent Variational Inference Algorithm

Coordinate Ascent Variational Inference (CAVI) is an optimization procedure for VI involving iterative updates to the proposal distribution to maximize the ELBO. CAVI uses the *mean-field approximation* expressed in Equation (5) to simplify the computation of update steps.

$$Q(\theta) = \prod_i Q(\theta_i) \tag{5}$$

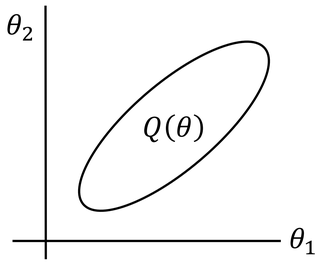

Figure 1: Unconstrained $Q(\theta)$

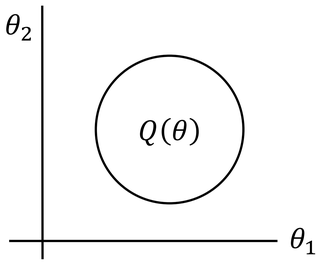

Figure 2: Mean-field approximation $Q(\theta)$

The mean-field approximation asserts that $Q(\theta)$ factorizes into independent $Q(\theta_i)$ as visualized for example $Q(\theta)$ in Figures 1 and 2. CAVI uses this approximation to compute an optimal update step for each $Q(\theta_i)$ while holding the remaining $\theta_{j\neq i}$ fixed. The update step for a single step of CAVI is expressed in Equation (6).

$$\underset{Q_i}{\text{argmax}}\, ELBO = \underset{Q_i}{\text{argmax}} \int_\theta Q(\theta) \log \frac{P(X, \theta)}{Q(\theta)} d\theta \tag{6}$$

We may compute the optimal Q_i by taking the functional derivative of the ELBO with respect to Q_i and setting to zero. Completing this computation leads to the following optimal Q_i:

$$Q^*(\theta_i) = \exp\left(\mathbb{E}_{Q_{j\neq i}}\left[P(\theta_i|X, \theta_{j\neq i})\right]\right) \tag{7}$$

CAVI takes iteratively updates each Q_i, continuously increasing the ELBO with each step. The update steps may be computed by plugging in the desired expressions for the likelihood and prior.

Example: Gaussian Mixture Model

A Gaussian mixture model attempts to describe a dataset as having been drawn IID from two or more Gaussian distributions. CAVI may be used to efficiently fit a Gaussian mixture model given prior beliefs about the model parameters. In this example, we will fit a Gaussian mixture model to synthetic two-dimensional data drawn from three normal distributions. The latent variables of the model are the cluster means, cluster covariance matrices, and the cluster weights. We compute the cluster responsibilities as well as model ELBO after each iteration.

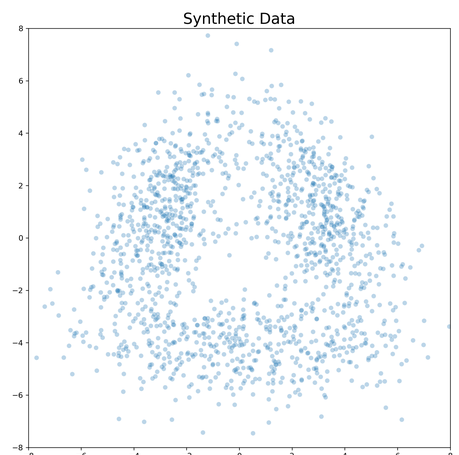

Figure 3: Synthetic data generated from three Gaussian distributions with additional added noise

Model Description

We adopt following model from the "full model" section given in [1]. In this model, ϕ gives the mixture weights, z_i gives the sample responsibilities, μ_k give the cluster means, and Σ_k give the cluster covariances.

Priors

$$\phi \sim Dirichlet(\alpha)$$
$$z_i \sim Categorical(\phi)$$
$$\mu_k \sim \mathcal{N}(\mu_0, I)$$
$$\Sigma_k \sim \mathcal{W}^{-1}(\Psi_0, \nu_0)$$

Data

$$X_i \sim \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$$

Variational Distribution

$$\phi \sim Dirichlet(\alpha_v)$$
$$z_i \sim Categorical(\phi_v)$$
$$\mu_k \sim \mathcal{N}(\mu_{vk}, I)$$
$$\Sigma_k \sim \mathcal{W}^{-1}(\Psi_v, \nu_v)$$

Gaussian Mixture Model Results

The model was fit using the BayesianGaussianMixture module from scikit-learn [7]. The model appeared to converge in about 80 iterations as demonstrated by the plot of the ELBO. The mixture clusters appeared to reflect the form of the data very closely. The below example used weak priors which enabled a close fit to the data, however, the user may specify more influential priors in other situations.

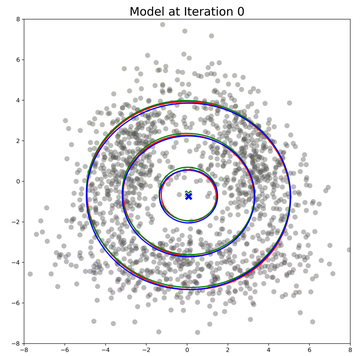

Figure 4: Visualization of model, iteration 0

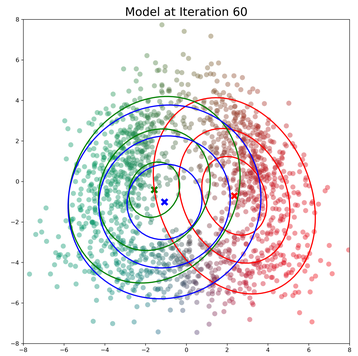

Figure 5: Visualization of model, iteration 60

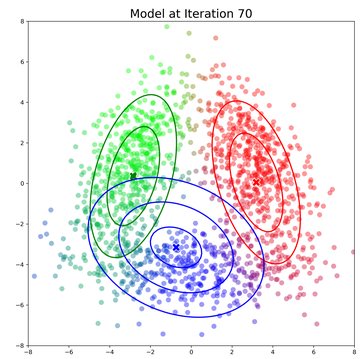

Figure 6: Visualization of model, iteration 70

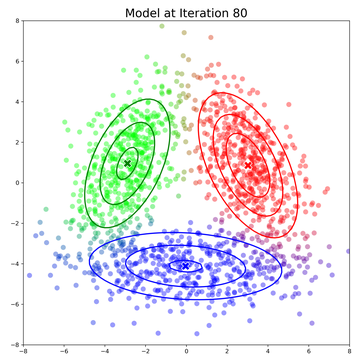

Figure 7: Visualization of model, iteration 80

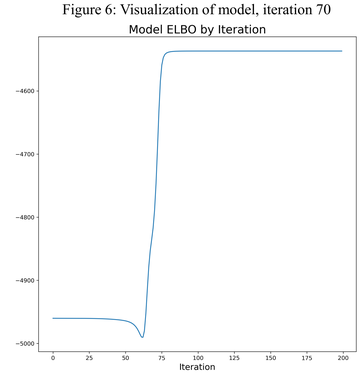

Figure 8: Plot of ELBO by training iteration

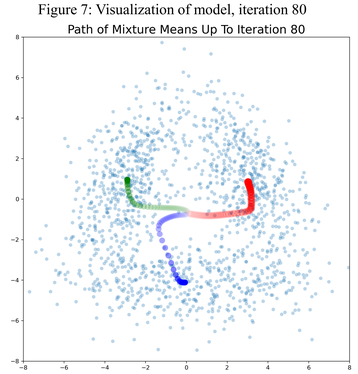

Figure 9: Path of variational means during training

References

This poster documents and extends work from my final project for STA 602: Bayesian Statistics taught by Professor Lin Lin. Each of the references below were used at some point during my original final project or during work for this poster.

[1] "2.1.3.2.1. Variational Gaussian Mixture Models." https://scikit-learn.org/0.15/modules/dp-derivation.html (accessed April 28, 2024).
[2] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdf.
[3] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, vol. 112, no. 518, pp. 859-877, 2017/04/03 2017, doi: 10.1080/01621459.2017.1285773.
[4] A. Ganguly and S. W. F. Earp, "An Introduction to Variational Inference," *ArXiv*, vol. abs/2108.13083, 2021.
[5] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," *CoRR*, vol. abs/1312.6114, 2013.
[6] A. Kushwaha. "Variational Inference: Gaussian Mixture model." https://ashkush.medium.com/variational-inference-gaussian-mixture-model-52595074247b (accessed 4/14, 2024).
[7] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[8] X. Y. S. Sia. "Coordinate Ascent Mean-field Variational Inference (Univariate Gaussian Example)." https://suzyahyah.github.io/bayesian%20inference/machine%20learning/2019/03/20/CAVI.html (accessed 4/14, 2024).
[9] C. Zhang, J. Bütepage, H. Kjellström, and S. Mandt, "Advances in Variational Inference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 2008-2026, 2017.