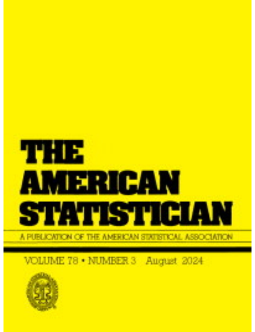

THE AMERICAN STATISTICIAN

A PUBLICATION OF THE AMERICAN STATISTICAL ASSOCIATION

VOLUME 78 • NUMBER 3 August 2024

Taylor & Francis
Taylor & Francis Group

The American Statistician

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/utas20

Introducing Variational Inference in Statistics and Data Science Curriculum

Vojtech Kejzlar & Jingchen Hu

To cite this article: Vojtech Kejzlar & Jingchen Hu (2024) Introducing Variational Inference in Statistics and Data Science Curriculum, The American Statistician, 78:3, 359-367, DOI: 10.1080/00031305.2023.2232006

To link to this article: https://doi.org/10.1080/00031305.2023.2232006

THE AMERICAN STATISTICIAN
2024, VOL. 78, NO. 3, 359–367: Teacher's Corner
https://doi.org/10.1080/00031305.2023.2232006

Taylor & Francis
Taylor & Francis Group

TEACHER'S CORNER

Introducing Variational Inference in Statistics and Data Science Curriculum

Vojtech Kejzlar[a] and Jingchen Hu[b]

[a]Department of Mathematics and Statistics, Skidmore College, Saratoga Springs, NY; [b]Department of Mathematics and Statistics, Vassar College, Poughkeepsie, NY

ABSTRACT
Probabilistic models such as logistic regression, Bayesian classification, neural networks, and models for natural language processing, are increasingly more present in both undergraduate and graduate statistics and data science curricula due to their wide range of applications. In this article, we present a one-week course module for students in advanced undergraduate and applied graduate courses on variational inference, a popular optimization-based approach for approximate inference with probabilistic models. Our proposed module is guided by active learning principles: In addition to lecture materials on variational inference, we provide an accompanying class activity, an R `shiny` app, and guided labs based on real data applications of logistic regression and clustering documents using Latent Dirichlet Allocation with R code. The main goal of our module is to expose students to a method that facilitates statistical modeling and inference with large datasets. Using our proposed module as a foundation, instructors can adopt and adapt it to introduce more realistic case studies and applications in data science, Bayesian statistics, multivariate analysis, and statistical machine learning courses.

ARTICLE HISTORY
Received January 2023
Accepted June 2023

KEYWORDS
Active learning; Bayesian statistics; Probabilistic models; Statistical computing; Variational inference

1. Introduction

With the recent and rapid expansion of both undergraduate and graduate curricula with offerings in data science, Bayesian statistics, multivariate data analysis, and statistical machine learning, probabilistic models and Bayesian methods have grown to become more popular (Schwab-McCoy, Baker, and Gasper 2021; Dogucu and Hu 2022). In many settings, a central task in applications of probabilistic models is the evaluation of posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ of m model parameters $\boldsymbol{\theta} \in \mathbb{R}^m$ $(m \geq 1)$ conditioned on the observed data $\boldsymbol{y} = (y_1, \ldots, y_n)$ provided by the Bayes' theorem

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{1}$$

Here, $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the sampling density given by the underlying probabilistic model for data, $p(\boldsymbol{\theta})$ is the prior density that represents our prior beliefs about $\boldsymbol{\theta}$ before seeing the data, and $p(\boldsymbol{y})$ is the marginal data distribution. The posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, however, has closed form only in a limited number of scenarios (e.g., conjugate priors) and therefore typically requires approximation. By far the most popular approximation methods are Markov chain Monte Carlo (MCMC) algorithms including Gibbs sampler, Metropolis, Metropolis-Hastings, and Hamiltonian Monte Carlo (Gelman et al. 2013), to name a few. See Albert and Hu (2020) for a review of these algorithms in undergraduate Bayesian courses. While useful for certain scenarios, these MCMC algorithms do not scale well with large datasets and can have a hard time approximating multimodal posteriors (Rudoy and Wolfe 2006; Bardenet, Doucet, and Holmes 2017). Such challenges therefore limit the applications of probabilistic models that can be discussed in the classroom and restrict students' exposure to more realistic case studies that include applying neural networks, pattern recognition, and natural language processing to massive datasets.

Variational inference is an alternative to the sampling-based approximation via MCMC that approximates a target density through optimization. Statisticians and computer scientists (starting with Peterson and Anderson 1987; Jordan et al. 1999; Blei, Kucukelbir, and McAuliffe 2017) have been using variational techniques in a variety of settings because these techniques tend to be faster and easier to scale to massive datasets. Despite its popularity among statistics and data science practitioners, variational inference is rarely discussed, especially in undergraduate courses, as it is believed to be a too advanced topic (Dogucu and Hu 2022). With this in mind, we have developed a one-week course module that serves as a gentle introduction to this topic. The goal is to help instructors to introduce variational techniques in their advanced undergraduate and applied graduate courses for more realistic case studies of probabilistic models. Our proposed one-week module is based on the best practices of active learning, which have been shown to improve student learning and engagement (Michael 2006; Freeman et al. 2014; Deslauriers et al. 2019). Our main guiding principle in designing the module is to involve students in the learning process by introducing student-centered class activities and labs. The guiding principle also includes assigning open-ended questions, focusing on problem-solving, providing appropriate scaffolding for activities, and creating opportunities to work collaboratively with peers.

CONTACT Vojtech Kejzlar ✉ *vkejzlar@skidmore.edu* Department of Mathematics and Statistics, Skidmore College, Saratoga Springs, NY.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TAS*.

Table 1. Outline of the one-week variational inference module.

	Content
First class	Lecture: Fundamentals of variational inference Class activity: Probabilistic model for count data with variational inference
Second class	Lab: Logistic regression/Document clustering

Our module is designed for students to gain a fundamental understanding and practical experience with variational inference over the course of two class meetings. During the first meeting, students are exposed to the fundamentals of variational inferences including the Kullback-Leibler divergence, evidence lower bound, gradient ascent, and coordinate ascent. Additionally, they gain their first hands-on experience by applying variational inference to a simple probabilistic model for count data. To encourage and empower instructors to adopt and adapt this variational inference module, we provide an accompanying in-class handout and an R `Shiny` app with details in the supplementary materials. During the second class meeting, students work on a guided R lab to apply variational inference to a realistic scenario. We offer two lab options for instructors to choose from depending on the course level and student background. For advanced undergraduate courses, we provide a case study of U.S. women labor participation with logistic regression model. For more advanced and self-motivated undergraduate students and applied graduate students, we present an application of variational inference to clustering documents with Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). See Table 1 for the breakdown of the module.

As for the audience, we believe that the module can be seamlessly integrated into any advanced undergraduate or applied graduate course in data science, Bayesian statistics, multivariate data analysis, and statistical machine learning that covers topics on clustering, classification, or text analysis. The prerequisites needed for the module are a basic understanding of statistical modeling, probability distributions, and elementary calculus.

The remainder of the article is organized as the following. In Section 2, we provide an overview of variational inference essentials that can be readily used as a basis for a lecture instruction. Section 3 presents a motivating example and the Gamma-Poisson model for count data that serves as the first hands-on class activity with variational inference. In Section 4, we offer realistic case studies for variational inference with implementation details in R, which can be used as a computing lab. We end the article in Section 5 with a few concluding remarks.

2. Lecture: Foundations of Variational Inference

In this section, we introduce concepts and definitions of variational inference in Section 2.1, discuss the choices of variational families in Section 2.2, and present details of ELBO optimization in Section 2.3. We also include recommendations of variational families and ELBO optimization strategies with pedagogical considerations for an advanced undergraduate and applied graduate audience. Instructors can design their lecture based on these materials tailored to their needs.

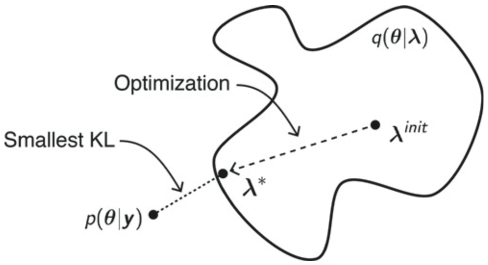
$q(\theta|\boldsymbol{\lambda})$
Optimization
Smallest KL
$\boldsymbol{\lambda}^{init}$
$\boldsymbol{\lambda}^*$
$p(\theta|\boldsymbol{y})$

Figure 1. Illustration of variational inference as the optimization-based approximation. The goal of variational inference is to find a member of the variational family that minimizes KL divergence with the target distribution.

2.1. Concepts and Definitions

The main idea behind variational inference is to approximate the target probability density $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ by a member of some relatively simple family of densities $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, indexed by the variational parameter $\boldsymbol{\lambda}$, over the space of model parameters $\boldsymbol{\theta}$. Note that $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)$ has m components of (potentially) varying dimensions. Variational approximation is done by finding the member of variational family that minimizes the Kullback-Leibler (KL) divergence of $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$:

$$q^* = \underset{q(\boldsymbol{\theta}|\boldsymbol{\lambda})}{\arg\min}\, KL(q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})||p(\boldsymbol{\theta} \mid \boldsymbol{y})), \tag{2}$$

with KL divergence being the expectation of the log ratio between the $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ with respect to $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$:

$$\begin{aligned} KL(q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})||p(\boldsymbol{\theta} \mid \boldsymbol{y})) &= \mathbb{E}_q\Big[\log \frac{q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})}{p(\boldsymbol{\theta} \mid \boldsymbol{y})}\Big] \\ &= \mathbb{E}_q\big[\log q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})\big] - \mathbb{E}_q\big[\log p(\boldsymbol{y}, \boldsymbol{\theta})\big] \\ &\quad + \log p(\boldsymbol{y}). \end{aligned} \tag{3}$$

The KL divergence measures how different is the probability distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ from $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ (Kullback and Leibler 1951). Note that while we use the KL divergence to measure the similarity between two densities, it is not a metric because the KL divergence is not symmetric and does not satisfy the triangle inequality. In fact, the order of $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ in (2) is deliberate as it leads to taking the expectation with respect to the variational distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$. One can naturally think of reversing the roles of $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ and $p(\boldsymbol{\theta} \mid \boldsymbol{y})$. However, this leads to a "different kind" of variational inference called *expectation propagation* (Minka (2001)), which loses computational efficiency of variational inference defined in (2).

In a nutshell, rather than sampling, variational inference approximates densities using optimization. See Figure 1 for a graphical illustration, that is, by finding the values of variational parameters from $\boldsymbol{\lambda}^{\text{init}}$ to $\boldsymbol{\lambda}^*$ through optimization which lead to a variational distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ that is close to the target posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ defined by the smallest KL divergence. Finding the optimal q^* is done in practice by maximizing an equivalent objective function, $\mathcal{L}(\boldsymbol{\lambda})$, the *evidence lower bound* (ELBO), because the KL divergence is intractable as it requires the evaluation of the marginal distribution $p(\boldsymbol{y})$:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\lambda}) = &\quad \mathbb{E}_q[\log p(\boldsymbol{y},\boldsymbol{\theta}) - \log q(\boldsymbol{\theta}|\boldsymbol{\lambda})]\\ = &\quad \underbrace{\mathbb{E}_q[\log p(\boldsymbol{y}|\boldsymbol{\theta})]}_{\text{Expected log-likelihood of data}}\\ - &\quad \underbrace{\mathrm{KL}(q(\boldsymbol{\theta}|\boldsymbol{\lambda})||p(\boldsymbol{\theta}))}_{\text{KL div. between the variational and prior densities}} \quad . \qquad (4)\end{aligned}$$

Starting with (3), one can derive the ELBO as the sum between the negative KL divergence of the variational density from the target density and the log of the marginal density $p(\boldsymbol{y})$. Since the term $\log p(\boldsymbol{y})$ is constant with respect to $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$, the objective functions in (3) and (4) are equivalent. Examining the ELBO also reveals the intuition behind variational inference. On the one hand, the first term in (4) encourages the variational approximation to place mass on parameter values that maximize the sampling density $p(\boldsymbol{y} \mid \boldsymbol{\theta})$. On the other hand, the second term in (4) prefers closeness of the variational density to the prior. Therefore, the ELBO shows a similar tension between the sampling density and the prior known in Bayesian inference.

2.2. Variational Families with Pedagogical Recommendations

We now move on to the implementation details of variational inference starting with the selection of the variational family $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$. This choice is crucial as it affects the complexity of optimization outlined in Section 2.1 as well as the quality of variational approximation.

2.2.1. Mean-Field Variational Family

By far the most popular is the *mean-field* variational family which assumes that all the unknown parameters are mutually independent, each approximated by its own univariate variational density:

$$q(\boldsymbol{\theta} \mid \boldsymbol{\lambda}) = \prod_{i=1}^{m} q(\theta_i \mid \boldsymbol{\lambda}_i). \qquad (5)$$

For example, a typical choice for real-valued parameters is the normal variational family $q(\theta \mid \mu, \sigma^2)$ and the log-normal or Gamma for nonnegative parameters. The main advantage of the mean-field family is in its simplicity as it requires only a minimum number of parameters to be estimated (no correlation parameters) and often leads to uncomplicated optimization. However, the mutually independent parameter assumption comes at a price because the mean-field family cannot capture relationships between model parameters. To illustrate the pitfalls of mean-field approximation, consider a simple case of a two-dimensional normal target density with highly correlated components. Figure 2 shows the optimal mean-field variational approximation given by the product of two normal densities. One can clearly see that the optimal variational densities match well with the means of the target density, but the marginal variances are underestimated. To further understand this common flaw of mean-field approximation, consider the definition of KL divergence in (3). The objective function penalizes more larger density in $q(\boldsymbol{\theta} \mid \boldsymbol{\lambda})$ in areas where $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ has low density than the opposite direction (recall that the expectation is taken with respect to the variational density).

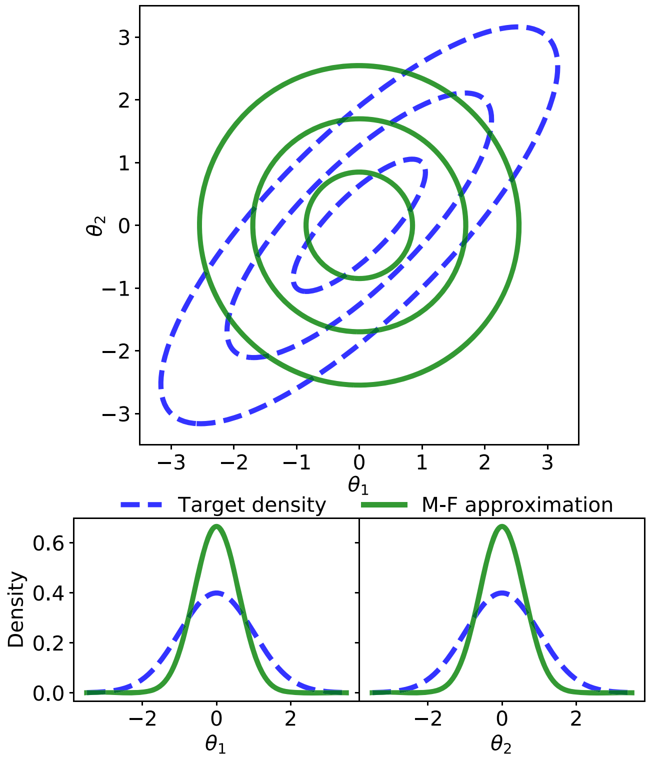

Figure 2. Mean-field variational approximation of a two-dimensional normal target density. The figure illustrates the common pitfall of the mean-field approximation in situations with correlated model parameters.

2.2.2. Recommendation for Instruction

It is worth noting that the development of new variational families which improves on the tradeoff between complexity and expressiveness of variational approximations has been a fruitful and active area of research. To keep the scope of the one-week variational inference module manageable to both the students and the instructors, we recommend solely focusing on the mean-field approximation. For interested students who want to explore further, we encourage the instructors to refer them to the recent work of Ambrogioni et al. (2021) that provides a detailed discussion on many state-of-the-art variational families and their associated implementation challenges.

2.3. ELBO Optimization with Pedagogical Recommendations

Besides the choice of variational family, another key implementation detail to address is the way in which we find the member of the variational family that maximizes the ELBO. Since this is a fairly general optimization problem, one can in principle use any optimization procedure. In the variational inference literature, the coordinate ascent and the gradient ascent procedures are the most prominent and widely used (Blei, Kucukelbir, and McAuliffe 2017).

2.3.1. Coordinate Ascent

The coordinate ascent approach is based on the simple idea that one can maximize ELBO, which is a multivariate function, by cyclically maximizing it along one direction at a time. Starting with initial values (denoted by superscript 0) of the m variational parameters $\boldsymbol{\lambda}^0$

$$\boldsymbol{\lambda}^0 = (\lambda_1^0, \dots, \lambda_m^0),$$

one obtains the $(k+1)$th updated value of variational parameters by iteratively solving

$$\lambda_i^{k+1} = \arg\max_x \mathcal{L}(\lambda_1^{k+1}, \dots, \lambda_{i-1}^{k+1}, x, \lambda_{i+1}^k, \dots, \lambda_m^k),$$

which can be accomplished without using gradients (Blei, Kucukelbir, and McAuliffe 2017).

2.3.2. Gradient Ascent

Variational inference via gradient ascent uses the standard iterative optimization algorithm based on the idea that the ELBO grows fastest in the direction of its gradient (Hoffman et al. 2013). In particular, the update of variational parameters $\boldsymbol{\lambda}$ at the $(k+1)$th iteration is given by

$$\boldsymbol{\lambda}^{k+1} \leftarrow \boldsymbol{\lambda}^k + \eta \times \nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda}^k),$$

where $\nabla_{\boldsymbol{\lambda}} \mathcal{L}(\boldsymbol{\lambda})$ is the ELBO gradient, and η is the step size which is also called the learning rate. The step size controls the rate at which one updates the variational parameters.

For both coordinate and gradient ascent, we typically declare convergence of variational parameters once the change in ELBO falls below some small threshold (Blei, Kucukelbir, and McAuliffe 2017).

2.3.3. Recommendation for Instruction

Our recommendation for this variational inference module is to take the route of gradient ascent. This pedagogical choice is guided by our combined experience of teaching statistical modeling, Bayesian statistics, and data science at various undergraduate levels to students with diverse statistical backgrounds. Our recommendation has also taken into account the pedagogical advantages and disadvantages of gradient ascent and coordinate ascent for the target audience: Variational inference via coordinate ascent, while conceptually straightforward, requires nontrivial and model-specific derivations which can easily obscure the overall goal of this one-week module to expand students' exposure to the state-of-the-art approximate inference for probabilistic models; gradient-based variational inference, in contrast, leads to a black-box optimization that does not require any model-specific derivations due to an extensive autodifferentiation capabilities of modern statistical software such as `RStan` (Stan Development Team 2022) and `Python` packages `PyTorch` (Paszke et al. 2019) and `TensorFlow` (Abadi et al. 2015), to name a few.

We believe that from an advanced undergraduate- and applied graduate-level pedagogical perspective, gradient descent reflects better the current data science pipeline and allows the instruction to be focused on conceptual understanding of variational inference rather than technical details. Of course, using gradient-based optimization requires the students to be familiar with partial derivatives. Such a prerequisite potentially restricts the audience for our module to a course with a multivariable calculus prerequisite. Nevertheless, we believe that an instructor with sufficient preparation can explain the basics behind gradient ascent to an audience with a minimal calculus background.

3. Class Activity: A Probabilistic Model for Count Data with Variational Inference

In this section, we provide a fully developed hands-on class activity with variational inference for count data. Starting with a motivating example in Section 3.1, we give an overview of the Gamma-Poisson model in Section 3.2, and discuss details of the variational inference of this model in Section 3.3, illustrated with an R `Shiny` app we have developed for instruction purpose. Instructors can adopt and adapt this class activity based on these materials tailored to their needs.

3.1. A Motivating Example

To illustrate how ELBO optimization leads to a good approximation of target posterior distribution, we consider Poisson sampling with a Gamma prior, which is a popular one-parameter model for count data (Gelman et al. 2013; Albert and Hu 2019; Johnson, Ott, and Dogucu 2022). To get started, we provide the following motivating example:

> Our task is to estimate the average number of active users of a popular massively multiplier online role-playing game (mmorpg) playing between the peak evening hours 7 pm and 10 pm. This information can help game developers in allocating server resources and optimizing user experience. To estimate the average number of active users, we will consider the counts (in thousands) of active players collected during the peak evening hours over a two-week period in the past month.

We have chosen the Gamma-Poisson model as the probabilistic model in this class activity for two reasons. First, the Gamma-Poisson model is relatively easy to understand for students with an elementary knowledge of probability distributions. Second, the Gamma is a conjugate prior for Poisson sampling which means that one can derive the exact posterior distribution (another Gamma) and check the fidelity of variational approximation by comparing to the analytical Gamma solution. The learning objective of this class activity is to get students familiarized with various aspects of variational inference presented in Section 2, such as ELBO and variational family, with a simple example. Afterwards, students are better prepared to move on to more realistic scenarios described in Section 4.

3.2. Overview of the Gamma-Poisson Model

We now provide an overview of the Gamma-Poisson model which can be readily turned into a class lecture. Suppose $\boldsymbol{y} = (y_1, \dots, y_n)$ represent the observed counts in n time intervals where the counts are independent, and each y_i follows a Poisson distribution with the same rate parameter $\theta > 0$. The joint probability mass function of $\boldsymbol{y} = (y_1, \dots, y_n)$ is

$$p(\boldsymbol{y} \mid \theta) = \prod_{i=1}^n p(y_i \mid \theta) \propto \theta^{\sum_{i=1}^n y_i} e^{-n\theta}. \tag{6}$$

The posterior distribution for the rate parameter θ is our inference target as θ represents the expected number of counts that occurs during the given time intervals. Note that the Poisson sampling relies on several assumptions about the sampling process: One assumes that the time interval is fixed, the counts occurring during different time intervals are independent, and the rate θ at which the counts occur is constant over time.

The Gamma-Poisson conjugacy states that if θ follows a Gamma prior distribution with shape and rate parameters α and β, it can be shown that the posterior distribution $p(\theta \mid \boldsymbol{y})$ will also have a Gamma density. Namely, if

$$\theta \sim \text{Gamma}(\alpha, \beta), \tag{7}$$

then

$$\theta \mid \boldsymbol{y} \sim \text{Gamma}(\alpha + \sum_{i=1}^{n} y_i, \beta + n). \tag{8}$$

In other words, given α, β, and $\boldsymbol{y}$, one can derive the analytical solution to the posterior of $p(\theta \mid \boldsymbol{y})$ and can subsequently sample from Gamma$(\alpha + \sum_{i=1}^{n} y_i, \beta + n)$ to get posterior samples of θ. While no approximation is needed, it serves as a good example of illustrating how variational inference works in such a setting and allows evaluations of the performance of variational inference.

3.3. *Variational Inference of the Gamma-Poisson Model*

Recall from Section 2 that variational inference approximates the (unknown) posterior distribution of a parameter by a simple family of distributions. In this Gamma-Poisson case, we will approximate the posterior distribution $p(\theta \mid \boldsymbol{y})$ by a log-normal distribution with mean μ and standard deviation σ:

$$q(\theta \mid \mu, \sigma) = \frac{1}{\theta\sigma\sqrt{2\pi}} e^{-\frac{(\ln\theta - \mu)^2}{2\sigma^2}}. \tag{9}$$

The log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. It is a popular variational family for nonnegative parameters because it can be expressed as a (continuously) transformed normal distribution, and therefore it is amenable to automatic differentiation. Automatic differentiation is a computation method for derivatives in computer programs that relies on the application of chain rule in differential calculus. It provides accurate and fast numerical derivative evaluations that leads to machine learning algorithms (such as variational inference) that do not require users to manually work out and code derivatives (Kucukelbir et al. 2017; Baydin et al. 2018).

In the supplementary materials, we provide an accompanying in-class handout and an R `Shiny` app based on the motivating scenario of mmorpg described in Section 3.1. The first two parts of the handout present the motivating example and the overview of the Gamma-Poisson model. In the third part of the handout, students carry out exact posterior inference for the unknown rate parameter θ using a small dataset of observed counts of mmorpg's active players. In the fourth and final part, students find variational approximation of $p(\theta \mid \boldsymbol{y})$ and check how well their approximation matches the true posterior distribution. Figure 3 shows the final variational approximation compared to the true Gamma(792, 100) posterior distribution from the handout example. We can see, on the one hand, the resulting log-normal(3.9, 0.04) distribution (the black dash line) that maximizes the ELBO visually overlaps with the true posterior (ELBO $= -42.52$, KL divergence < 0.001). On the other hand, another member of the variational family, the log-normal(3.7, 0.05) distribution (the blue dot-dash line; with ELBO $= -57.55$ and KL divergence $=$ 15.085), clearly differs from the target. This example illustrates the good performance of variational inference through optimization for the Gamma-Poisson count model.

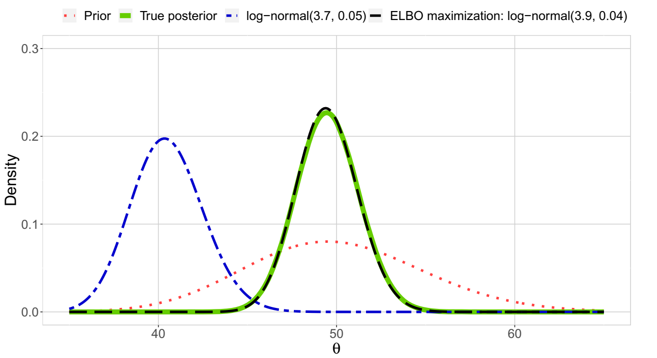

Figure 3. Variational approximation based on the motivating scenario of mmorpg's player activity. The true Gamma(792, 100) posterior and the prior Gamma(100,2) distributions are included.

The design of this class activity is guided by the active-learning principles listed in Section 1 and the goal is to give students their first hands-on experience with variational inference without the need of coding. Specifically, we include open-ended questions that focus on problem-solving and create opportunities for students to collaborate with peers. Moreover, the accompanying R `Shiny` app provides appropriate and sufficient scaffolding so that students can concentrate on conceptual understanding instead of the technical details, which follows our pedagogical recommendations in Section 2.

We now turn to two guided `R` labs to illustrate the use of variational inference for more realistic case studies of logistic regression and document clustering.

4. Labs: Logistic Regression and Document Clustering

In what follows, we provide two alternatives for the lab portion of the proposed module. Section 4.1 outlines a case study of U.S. women labor participation with logistic regression model aimed for an advanced undergraduate audience. Section 4.1 applies variational inference on document clustering of a collection of Associate Press newspaper articles targeted for more advanced and self-motivated undergraduate students and students in applied graduate courses.

4.1. *Logistic Regression*

Logistic regression model is a popular supervised learning algorithm for binary classification due to its interpretability, solid predictive performance, and intuitive connection to the standard linear regression (James et al. 2013). Despites its popularity, logistic regression, and its Bayesian version in particular, poses computational and statistical challenges in scenarios with large and high-dimensional data (Genkin, Lewis, and Madigan 2007). For these reasons, we believe that a logistic regression is a suitable lab for an advanced (or an intermediate) undergraduate audience that can demonstrate the benefits of variational inference with a relatively low barrier from the statistical methodology point of view.

In Section 4.1.1, we briefly introduce the logistic regression model. In Section 4.1.2, we present a case study of U.S. women labor participation analysis where variational inference is implemented by the `cmdstanr` R package. We mainly focus on the interpretation of results and discuss pedagogical considerations and leave the details of the guided lab assignment with R code in the supplementary materials.

4.1.1. Overview of Logistic Regression

The logistic regression model assumes that a binary response y_i follows a Bernoulli distribution with probability of success p_i:

$$y_i \mid p_i \sim \text{Bernoulli}(p_i).$$

To relate a single predictor x_i to the response y_i, logistic regression typically considers the natural logarithm of odds $p_i/(1-p_i)$ (also known as logit) to be a linear function of the predictor variable x_i:

$$\text{logit}(p_i) = \ln\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta x_i, \tag{10}$$

with α and β being regression coefficients. Note that it is a bit more challenging to interpret the coefficients in the logistic regression than in standard linear regression as α and β are directly related to the log odds $p_i/(1-p_i)$, instead of p_i. For example, e^{α} is the odds when the value of predictor x_i is 0, whereas the quantity e^{β} refers to the change in odds per unit increase in x_i.

Lastly, by rearranging the terms in (10), one can express the probability of success p_i as

$$p_i = \frac{e^{\alpha+\beta x_i}}{1+e^{\alpha+\beta x_i}}.$$

In the Bayesian framework, one proceeds to prior specification of regression coefficients (α, β) and posterior inference through MCMC. For illustration, we consider independent normal priors for the regression coefficients $\alpha \sim \text{Normal}(\mu_0, \sigma_0)$ and $\beta \sim \text{Normal}(\mu_1, \sigma_1)$, where (μ_0, μ_1) and (σ_0, σ_1) are the prior means and standard deviations for the regression coefficients, respectively.

4.1.2. Predicting Labor Participation

To apply variational inference in the context of logistic regression, we present a case study of predicting U.S. women labor participation. To do so, we consider a sample from the University of Michigan Panel Study of Income Dynamics (PSID), the longest running longitudinal household survey in the world. The survey dates back to 1968 and contains information on over 18,000 individuals living in 5000 families in the United States. The survey of these individuals and their descendants has been collected continuously and includes data on income, wealth, employment status, health, marriage, and hundreds of other variables. Our interest is in analyzing a PSID sample of 753 observations from 1976 (Mroz 1987). The PSID 1976 survey is particularly interesting since it interviewed wives in households directly in the previous year. This PISD sample contains two variables: Family income exclusive of wife's income (in $1000) and wife's labor participation (yes or no). The goal of the lab is predicting a wife's labor participation status (response variable y_i) from the family income exclusive of her income (predictor variable x_i) using logistic regression. We refer interested readers to Albert and Hu (2019) Section 11.4 for an in-depth illustration of Bayesian logistic regression applied to the same prediction task.

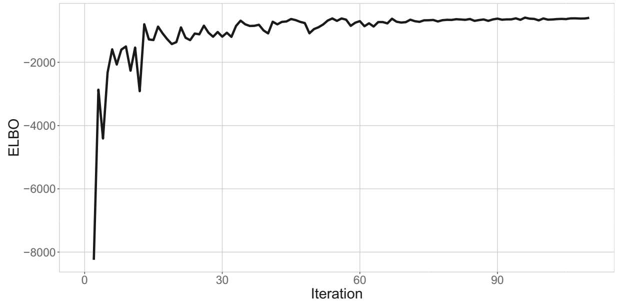

Figure 4. The evolution of ELBO for the logistic regression model based on a PSID sample of 753 observations from 1976.

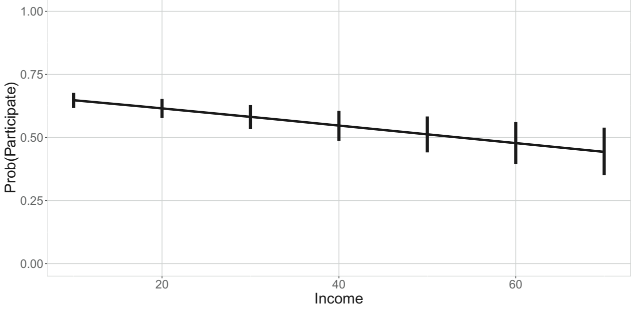

Figure 5. Posterior interval estimates for the probability of labor participation of a married woman who has a family income exclusive of her own income.

Figure 4 shows the evolution of ELBO for the logistic regression model which converged after 80 iterations of the gradient ascent algorithm described in Section 2.3. We recommend running the algorithm repeatedly (i.e., 2–3 times) with a different random seed in the classroom and discussing the dependency of variational inference on initial values of variational parameters which can occur in practice. To highlight the computational benefits of variational inference, we also propose generating 50 replicates of the PSID sample (37,650 observations in total) and comparing the speed of convergence of variational approximation and MCMC approximation. The details of this exercise are provided in the supplementary materials.

Figure 5 demonstrates one of the potential insights of the PSID survey data analysis, which is a series of posterior interval estimates for the probability of labor participation of a married woman who has a family income exclusive of her own income ranging from $10,000 to $70,000 with $10,000 increments. One can see that in 1976, the likelihood of labor participation decreased with increasing family income exclusive of wife's income.

4.2. Document Clustering

Among the many models approximated by variational inference techniques, Latent Dirichlet Allocation (LDA) might be one of the most popular (Blei, Ng, and Jordan 2003). LDA is a mixed-membership clustering model, commonly used for document clustering. Specifically, LDA models each document to have a mixture of topics, where each word in the document is drawn from a topic based on the mixing proportions (Stan Development Team 2022). While the LDA model is relatively easy and straightforward to follow, using conventional MCMC estimation techniques has proven to be too computationally demanding due to the large number of parameters involved. Therefore, researchers and practitioners turn to variational inference techniques when using LDA for document clustering (Blei, Ng, and Jordan 2003).

In Section 4.2.1, we briefly introduce the LDA model following the presentation in Stan Development Team (2022). In Section 4.2.2, we present an LDA application to a collection of Associate Press newspaper articles where variational inference is implemented by the `cmdstanr` R package. For brevity, we focus on the interpretation of results and discuss pedagogical considerations and leave a `Stan` script for the LDA model and the details of the guided lab assignment with `R` code in the supplementary materials.

4.2.1. Overview of the LDA Model

The LDA model considers K topics for M documents made up of words drawn from a vocabulary of V distinct words. For a document m, a topic distribution $\boldsymbol{\theta_m}$ over K topics is drawn from a Dirichlet distribution,

$$\boldsymbol{\theta}_m \sim \text{Dirichlet}(\boldsymbol{\alpha}), \tag{11}$$

where $\sum_{k=1}^{K} \theta_{m,k} = 1$ $(0 \leq \theta_{m,k} \leq 1)$ and $\boldsymbol{\alpha}$ is a vector of length K with positive values.

Each of the N_m words $\{w_{m,1}, \ldots, w_{m,N_m}\}$ in document m is then generated independently conditional on $\boldsymbol{\theta}_m$. To do so, first, the topic $z_{m,n}$ for word $w_{m,n}$ in document m is drawn from

$$z_{m,n} \sim \text{categorical}(\boldsymbol{\theta}_m), \tag{12}$$

where $\boldsymbol{\theta}_m$ is the document-specific topic-distribution defined in (11).

Next, the word $w_{m,n}$ in document m is drawn from

$$w_{m,n} \sim \text{categorical}(\boldsymbol{\phi}_{z[m,n]}), \tag{13}$$

which is the word distribution for topic $z_{m,n}$. Note that $z[m,n]$ in (13) refers to $z_{m,n}$.

Lastly, a Dirichlet prior is given to distributions $\boldsymbol{\phi}_k$ over words for topic k as

$$\boldsymbol{\phi}_k \sim \text{Dirichlet}(\boldsymbol{\beta}), \tag{14}$$

where $\boldsymbol{\beta}$ is the prior a vector of length V (i.e., the total number of words) with positive values. Figure 6 shows a graphical model representation of LDA.

4.2.2. Clustering of Associated Press Newspaper Articles

As a realistic application of variational inference, we consider a collection of 2246 Associated Press newspaper articles to be clustered using the LDA model. The dataset is (conveniently) part of the `topicmodels` R package. We believe this dataset is well suited to demonstrate the capabilities of variational inference in the classroom as it is too large for the MCMC approximation to be feasible but small enough for the variational inference to take just a few minutes to converge. For brevity, we highlight the results based on a two-topic LDA model (i.e., $K = 2$) and leave the details to the guided lab in the supplementary materials. The number of topics is set to 2 for demonstration purposes and simplicity of interpretations. Comparing LDA with a different number of topics is often done with metrics such as semantic coherence or held-out data likelihood (Mimno et al. 2011). While such a comparison is beyond the scope of this lab, interested students are encouraged to explore while being mentored by the instructors.

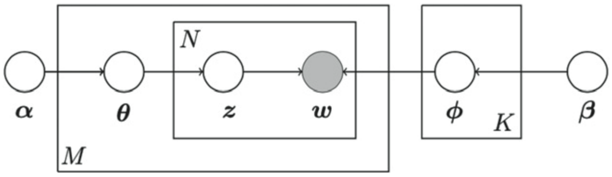

Figure 6. Graphical model representation of LDA. The largest box represents the documents. On the left, the inner box represents the topics and words within each document. On the right, the box represents the topics.

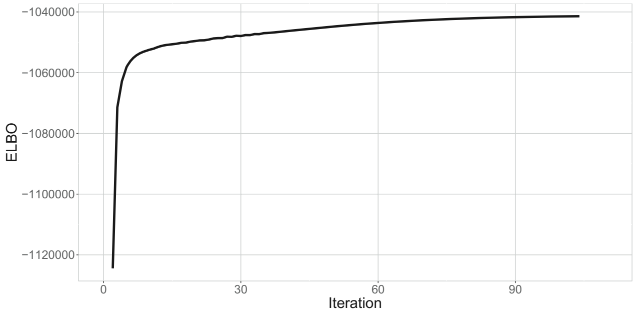

Figure 7. The evolution of ELBO for the two-topic LDA model based on 2246 Associated Press newspaper articles.

Figure 7 shows the evolution of ELBO for the two-topic LDA model which converged after a little bit over 100 iterations of the gradient ascent algorithm described in Section 2.3. On a standard laptop computer, this typically takes between 5 and 10 min depending on the CPU speed. Similar to the U.S. labor participation case study, we recommend running the algorithm repeatedly (i.e., 2–3 times) with a different random seed in the classroom and discussing the dependency of variational inference on initial values of variational parameters which can occur in practice.

Figures 8 and 9 are examples of graphical displays of the topics that were extracted from the collection of articles based with the LDA. In particular, Figure 8 shows the 10 most common words for each topic; that is, the parts of distribution $\boldsymbol{\phi}_k$, for $k \in \{1,2\}$, with the largest mass. Figure 9 displays similar information for the 20 most common words for each topic in the form of a word cloud. The most common words in topic 1 include *people, government, president, police,* and *state*, suggesting that this topic may represent political news. In contrast, the most common words in topic 2 include *percent, billion, million, market, American,* and *states*, hinting that this topic may represent news about the U.S. economy.

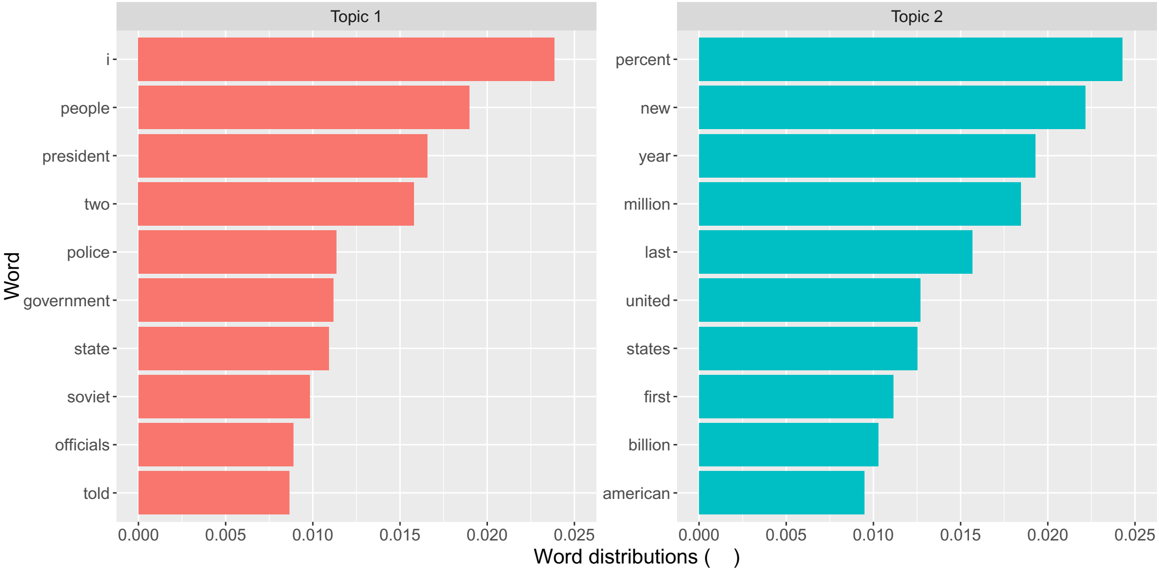

Figure 8. Word distributions based on the two-topic LDA model. The 10 most common words are displayed.

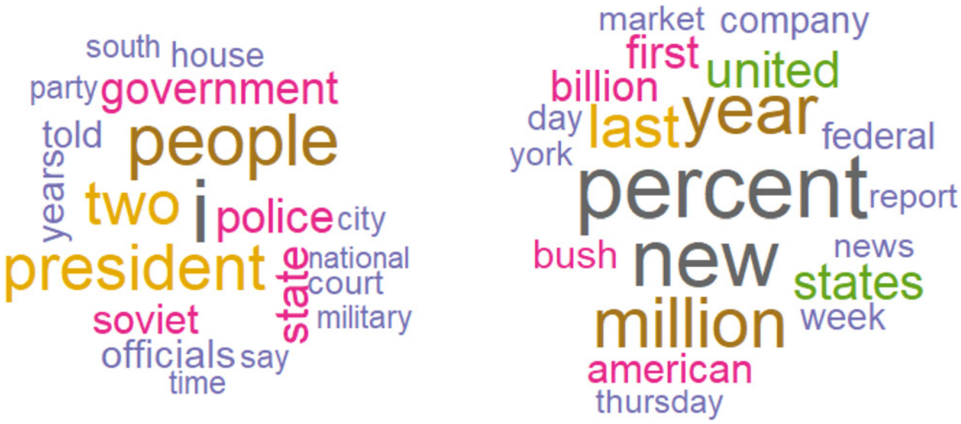

Figure 9. World clouds consisting of the 20 most common words for each of the two topics extracted by the LDA.

5. Concluding Remarks

In this article, we present a newly-developed one-week course module that exposes students in advanced undergraduate and applied graduate courses to approximation via variational inference. The proposed module is self-contained in the sense that it encourages and empowers potential instructors to adopt and adapt the module as we provide an overview of variational inference, an active-learning-based class activity with an R `Shiny` app, and guided labs based on a realistic application with R code (see the supplementary materials or *https://github.com/kejzlarv/variational_inference_module*). Its design is rooted in the best practices of active learning that have been demonstrated to improve student learning and engagement.

The module can be integrated into any advanced undergraduate or applied graduate course where students learn probabilistic models (including logistic regression, Bayesian classifiers, neural networks, or models for natural language processing), such as Bayesian statistics, multivariate data analysis, and data science courses. The applications discussed in these courses are typically limited to scenarios with relatively small datasets, since the required use of MCMC does not scale well to large datasets. Given the popularity and scalability of variational inference, we hope that instructors adopting and adapting this module will be able to integrate more realistic and fun case studies in their classrooms. Moreover, the references and further readings provided in this article are readily available resources for a deeper dive of variational inference by interested students with appropriate mentoring by their instructors.

Supplementary Materials

The supplementary files for this article include the following: (a) Details of the class activity on probabilistic model for count data with variational inference; (b) The manual and the `R shiny` app we have developed for the module; (c) Details of the guided `R` logistic regression lab with U.S. women labor participation sample data; and (d) Details of the guided `R` lab of the LDA application to a sample of the Associated Press newspaper articles with variational inference.

Disclosure Statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding

Jingchen Hu acknowledges the support from the National Science Foundation grant DUE-2215920.

ORCID

Vojtech Kejzlar https://orcid.org/0000-0002-1001-011X
Jingchen Hu https://orcid.org/0000-0002-4283-181X

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015), "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," Software available from tensorflow.org. Available at *https://www.tensorflow.org/*. [362]

Albert, J., and Hu, J. (2019), *Probability and Bayesian Modeling* (1st ed.), Boca Raton, FL: Chapman and Hall/CRC. [362,364]

——— (2020), "Bayesian Computing in the Undergraduate Statistics Curriculum," *Journal of Statistics Education*, 28, 236–247. [359]

Ambrogioni, L., Lin, K., Fertig, E., Vikram, S., Hinne, M., Moore, D., and van Gerven, M. (2021), "Automatic Structured Variational Inference," in *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, vol. 130 of Proceedings of Machine Learning Research, PMLR, eds. A. Banerjee and K. Fukumizu, pp. 676–684. Available at *https://proceedings.mlr.press/v130/ambrogioni21a.html* [361]

Bardenet, R., Doucet, A., and Holmes, C. (2017), "On Markov Chain Monte Carlo Methods for Tall Data," *Journal of Machine Learning Research*, 18, 1515–1557. [359]

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018), "Automatic Differentiation in Machine Learning: A Survey," *Journal of Machine Learning Research*, 18, 1–43. Available at *http://jmlr.org/papers/v18/17-468.html* [363]

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017), "Variational Inference: A Review for Statisticians," *Journal of the American Statistical Association*, 112, 859–877. DOI:10.1080/01621459.2017.1285773 [359,361,362]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [360,365]

Deslauriers, L., McCarty, L. S., Miller, K., Callaghan, K. and Kestin, G. (2019), "Measuring Actual Learning Versus Feeling of Learning in Response to Being Actively Engaged in the Classroom," *Proceedings of the National Academy of Sciences*, 116, 19251–19257. Available at *https://www.pnas.org/doi/abs/10.1073/pnas.1821936116* [359]

Dogucu, M., and Hu, J. (2022), "The Current State of Undergraduate Bayesian Education and Recommendations for the Future," *The American Statistician*, 76, 405–413. DOI:10.1080/00031305.2022.2089232 [359]

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., and Wenderoth, M. P. (2014), "Active Learning Increases Student Performance in Science, Engineering, and Mathematics," *Proceedings of the National Academy of Sciences*, 111, 8410–8415. DOI:10.1073/pnas.1319030111 [359]

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013), *Bayesian Data Analysis* (3rd ed.), Boca Raton, FL: CRC Press. Available at *https://books.google.com/books?id=ZXL6AQAAQBAJ* [359,362]

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304. DOI:10.1198/004017007000000245 [363]

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. Available at *http://jmlr.org/papers/v14/hoffman13a.html* [362]

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013), *An Introduction to Statistical Learning: with Applications in R*, Springer Texts in Statistics, New York: Springer. Available at *https://books.google.com/books?id=qcI_AAAAQBAJ* [363]

Johnson, A. A., Ott, M., and Dogucu, M. (2022), *Bayes Rules! An Introduction to Applied Bayesian Modeling* (1st ed.), Boca Raton, FL: Chapman and Hall/CRC. [362]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [359]

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic Differentiation Variational Inference," *Journal of Machine Learning Research*, 18, 1–45. Available at *http://jmlr.org/papers/v18/16-107.html* [363]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86. DOI:10.1214/aoms/1177729694 [360]

Michael, J. (2006), "Where's the Evidence that Active Learning Works?," *Advances in Physiology Education*, 30, 159–167. DOI:10.1152/advan.00053.2006 [359]

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011), "Optimizing Semantic Coherence in Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Association for Computational Linguistics, USA, pp. 262–272. [365]

Minka, T. P. (2001), "Expectation Propagation for Approximate Bayesian Inference," in *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 362–369. [360]

Mroz, T. A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799. [364]

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019), "Pytorch: An Imperative Style, High-Performance Deep Learning Library," *in Advances in Neural Information Processing Systems 32*, Curran Associates, Inc., pp. 8024–8035. Available at *http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf* [362]

Peterson, C., and Anderson, J. R. (1987), "A Mean Field Theory Learning Algorithm for Neural Networks," *Complex Systems*, 1, 995–1019. [359]

Rudoy, D., and Wolfe, P. J. (2006), "Monte Carlo Methods for Multi-Modal Distributions," in *2006 Fortieth Asilomar Conference on Signals, Systems and Computers*, pp. 2019–2023. [359]

Schwab-McCoy, A., Baker, C. M., and Gasper, R. E. (2021), "Data Science in 2020: Computing, Curricula, and Challenges for the Next 10 Years," *Journal of Statistics and Data Science Education*, 29, S40–S50. DOI:10.1080/10691898.2020.1851159 [359]

Stan Development Team (2022), *Stan Modeling Language User's Guide and Reference Manual, Version 2.31*. Available at *http://mc-stan.org/* [362,365]