entropy

MDPI

Article

Objective Bayesian Inference in Probit Models with Intrinsic Priors Using Variational Approximations

Ang Li [†,‡], **Luis Pericchi** [*,†,‡] **and Kun Wang** [†,‡]

Río Piedras Campus, University of Puerto Rico, 00925 San Juan, Puerto Rico; ang.li@upr.edu (A.L.); kun.wang@upr.edu (K.W.)

[*] Correspondence: luis.pericchi@upr.edu

[†] 14 Ave. Universidad Ste. 1401, San Juan, PR 00925, USA.

[‡] These authors contributed equally to this work.

Received: 3 March 2020; Accepted: 26 April 2020; Published: 30 April 2020

Abstract: There is not much literature on objective Bayesian analysis for binary classification problems, especially for intrinsic prior related methods. On the other hand, variational inference methods have been employed to solve classification problems using probit regression and logistic regression with normal priors. In this article, we propose to apply the variational approximation on probit regression models with intrinsic prior. We review the mean-field variational method and the procedure of developing intrinsic prior for the probit regression model. We then present our work on implementing the variational Bayesian probit regression model using intrinsic prior. Publicly available data from the world's largest peer-to-peer lending platform, LendingClub, will be used to illustrate how model output uncertainties are addressed through the framework we proposed. With LendingClub data, the target variable is the final status of a loan, either charged-off or fully paid. Investors may very well be interested in how predictive features like FICO, amount financed, income, etc. may affect the final loan status.

Keywords: objective Bayesian inference; intrinsic prior; variational inference; binary probit regression; mean-field approximation

1. Introduction

There is not much literature on objective Bayesian analysis for binary classification problems, especially for intrinsic prior related methods. By far, only two articles have explored intrinsic prior related methods on classification problems. Reference [1] implements integral priors into the generalized linear models with various link functions. In addition, reference [2] considers intrinsic priors for probit models. On the other hand, variational inference methods have been employed to solve classification problem with logistic regression ([3]) and probit regression ([4,5]) with normal priors. Variational approximation methods have been reviewed in [6,7], and more recently [8].

In this article, we propose to apply variational approximations on probit regression models with intrinsic priors. In Section 4, we review the mean-field variational method that will be used in this article. In Section 3, procedures for developing intrinsic priors for probit models will be introduced following [2]. Our work is presented in Section 5. Our motivations for combining intrinsic prior methodology and variational inference is as following

- Avoiding manually set ad hoc plugin priors by automatically generating a family of non-informative priors that are less sensible.
- Reference [1,2] do not consider inference of posterior distributions of parameters. Their focus is on model comparison. Although the development of intrinsic priors itself comes from a model

selection background, we thought it would be interesting to apply intrinsic priors on inference problems. In fact, some recently developed priors that proposed to solve inference or estimation problems turned out to be also intrinsic priors. For example, the Scaled Beta2 prior [9] and the Matrix-*F* prior [10].

- Intrinsic priors concentrate probability near the null hypothesis, a condition that is widely accepted and should be required of a prior for testing a hypothesis.
- Also, intrinsic priors have flat tails that prevents finite sample inconsistency [11].
- For inference problems with large data set, variational approximation methods are much faster than MCMC-based methods.

As for model comparison, due to the fact that the output of variational inference methods cannot be employed directly to compare models, we propose in Section 5.3 to simply make use of the variational approximation of the posterior distribution as an importance function and get the Monte Carlo estimated marginal likelihood by importance sampling for model comparison.

2. Background and Development of Intrinsic Prior Methodology

2.1. Bayes Factor

The Bayesian framework of model selection coherently involves the use of probability to express all uncertainty in the choice of model, including uncertainty about the unknown parameters of a model. Suppose that models $M_1, M_2, ..., M_q$ are under consideration. We shall assume that the observed data $\mathbf{x} = (x_1, x_2, ..., x_n)$ is generated from one of these models but we do not know which one it is. We express our uncertainty through prior probability $P(M_j), j = 1, 2, ..., q$. Under model M_i, $\mathbf{x}$ has density $f_i(\mathbf{x}|\boldsymbol{\theta}_i, M_i)$, where $\boldsymbol{\theta}_i$ are unknown model parameters, and the prior distribution for $\boldsymbol{\theta}_i$ is $\pi_i(\boldsymbol{\theta}_i|M_i)$. Given observed data and prior probabilities, we can then evaluate the posterior probability of M_i using Bayes' rule

$$P(M_i|\mathbf{x}) = \frac{p_i(\mathbf{x}|M_i)P(M_i)}{\sum_{j=1}^{q} p_j(\mathbf{x}|M_j)P(M_j)}, \tag{1}$$

where

$$p_i(\mathbf{x}|M_i) = \int f_i(\mathbf{x}|\boldsymbol{\theta}_i, M_i)\pi_i(\boldsymbol{\theta}_i|M_i)d\boldsymbol{\theta}_i \tag{2}$$

is the marginal likelihood of $\mathbf{x}$ under M_i, also called the evidence for M_i [12]. A common choice of prior model probabilities is $P(M_j) = \frac{1}{q}$, so that each model has the same initial probability. However, there are other alternatives of assigning probabilities to correct for multiple comparison (See [13]). From (1), the posterior odds are therefore the prior odds multiplied by the Bayes factor

$$\frac{P(M_j|\mathbf{x})}{P(M_i|\mathbf{x})} = \frac{P(M_j)p_j(\mathbf{x})}{P(M_i)p_i(\mathbf{x})} = \frac{P(M_j)}{P(M_i)} \times B_{ji}. \tag{3}$$

where the Bayes factor of M_j to M_i is defined by

$$B_{ji} = \frac{p_j(\mathbf{x})}{p_i(\mathbf{x})} = \frac{\int f_j(\mathbf{x}|\boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}. \tag{4}$$

Here we omit the dependence on models M_j, M_i to keep the notation simple. The marginal likelihood, $p_i(\mathbf{x})$ expresses the preference shown by the observed data for different models. When $B_{ji} > 1$, the data favor M_j over M_i, and when $B_{ji} < 1$ the data favor M_i over M_j. A scale for interpretation of B_{ji} is given by [14].

2.2. Motivation and Development of Intrinsic Prior

Computing B_{ji} requires specification of $\pi_i(\boldsymbol{\theta}_i)$ and $\pi_j(\boldsymbol{\theta}_j)$. Often in Bayesian analysis, when prior information is weak, one can use non-informative (or default) priors $\pi_i^N(\boldsymbol{\theta}_i)$. Common choices for non-informative priors are the uniform prior, $\pi_i^U(\boldsymbol{\theta}_i) \propto 1$; the Jeffreys prior, $\pi_i^J(\boldsymbol{\theta}_i) \propto \left[\det(\mathbf{I}_i(\boldsymbol{\theta}_i))\right]^{1/2}$ where $\mathbf{I}_i(\boldsymbol{\theta}_i)$ is the expected Fisher information matrix corresponding to M_i.

Using any of the π_i^N in (4) would yield

$$B_{ji}^N = \frac{p_j^N(\mathbf{x})}{p_i^N(\mathbf{x})} = \frac{\int f_j(\mathbf{x}|\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j)d\boldsymbol{\theta}_j}{\int f_i(\mathbf{x}|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i}. \tag{5}$$

The difficulty with (5) is that π_i^N are typically improper and hence are defined only up to an unspecified constant c_i. So B_{ji}^N is defined only up to the ratio c_j/c_i of two unspecified constants.

An attempt to circumvent the ill definition of the Bayes factors for improper non-informative priors is the intrinsic Bayes factor introduced by [15], which is a modification of a partial Bayes factor [16]. To define the intrinsic Bayes factor we consider the set of subsamples $\mathbf{x}(l)$ of the data $\mathbf{x}$ of minimal size l such that $0 < p_i^N(\mathbf{x}(l)) < \infty$. These subsamples are called training samples (not to be confused with training sample in machine learning). In addition, there is a total number of L such subsamples.

The main idea here is that training sample $\mathbf{x}(l)$ will be used to convert the improper $\pi_i^N(\boldsymbol{\theta}_i)$ to proper posterior

$$\pi_i^N(\boldsymbol{\theta}_i|\mathbf{x}(l)) = \frac{f_i(\mathbf{x}(l)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)}{p_i^N(\mathbf{x}(l))} \tag{6}$$

where $p_i^N(\mathbf{x}(l)) = \int f_i(\mathbf{x}(l)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$. Then, the Bayes factor for the remaining of the data $\mathbf{x}(n-l)$, where $\mathbf{x}(l) \cup \mathbf{x}(n-l) = \mathbf{x}$, using $\pi_i^N(\boldsymbol{\theta}_i|\mathbf{x}(l))$ as prior is called a "partial" Bayes factor,

$$B_{ji}^N(\mathbf{x}(n-l)|\mathbf{x}(l)) = \frac{\int f_j(\mathbf{x}(n-l)|\boldsymbol{\theta}_j)\pi_j^N(\boldsymbol{\theta}_j|\mathbf{x}(l))d\boldsymbol{\theta}_j}{\int f_i(\mathbf{x}(n-l)|\boldsymbol{\theta}_i)\pi_i^N(\boldsymbol{\theta}_i|\mathbf{x}(l))d\boldsymbol{\theta}_i} \tag{7}$$

This partial Bayes factor is a well-defined Bayes factor, and can be written as $B_{ji}^N(\mathbf{x}(n-l)|\mathbf{x}(l)) = B_{ji}^N(\mathbf{x})B_{ij}(\mathbf{x}(l))$, where $B_{ji}^N(\mathbf{x}) = \frac{p_j^N(\mathbf{x})}{p_i^N(\mathbf{x})}$ and $B_{ij}(\mathbf{x}(l)) = \frac{p_i^N(\mathbf{x}(l))}{p_j^N(\mathbf{x}(l))}$. Clearly, $B_{ji}^N(\mathbf{x}(n-l)|\mathbf{x}(l))$ will depend on the choice of the training samples $\mathbf{x}(l)$. To eliminate this arbitrariness and increase stability, reference [15] suggests averaging over all training samples and obtained the arithmetic intrinsic Bayes factor (AIBF)

$$B_{ji}^{AIBF}(\mathbf{x}) = B_{ji}^N(\mathbf{x})\frac{1}{L}\sum_{l=1}^{L} B_{ij}^N(\mathbf{x}(l)). \tag{8}$$

The strongest justification of the arithmetic IBF is its asymptotic equivalence with a proper Bayes factor arising from *Intrinsic priors*. These intrinsic priors were identified through an asymptotic analysis (see [15]). For the case where M_i is nested in M_j, it can be shown that the intrinsic priors are given by

$$\pi_i^I(\boldsymbol{\theta}_i) = \pi_i^N(\boldsymbol{\theta}_i) \text{ and } \pi_j^I(\boldsymbol{\theta}_j) = \pi_j^N(\boldsymbol{\theta}_j)E_{M_j}\left[\frac{m_i^N(\mathbf{x}(l))}{m_j^N(\mathbf{x}(l))}|\boldsymbol{\theta}_j\right]. \tag{9}$$

3. Objective Bayesian Probit Regression Models

3.1. Bayesian Probit Model and the Use of Auxiliary Variables

Consider a sample $\mathbf{y} = (y_1, ..., y_n)$, where $Y_i, i = 1, ..., n$, is a $0 - 1$ random variable such that under model M_j, it follows a probit regression model with a $j + 1$-dimensional vector of covariates x_i, where $j \leq p$. Here, p is the total number of covariate variables under our consideration. In addition, this probit model M_j has the form

$$Y_i|\beta_0, ..., \beta_j, M_j \sim \text{Bernoulli}(\Phi(\beta_0 x_{0i} + \beta_1 x_{1i} + ... + \beta_j x_{ji})), \qquad 1 \leq i \leq n, \tag{10}$$

where Φ denotes the standard normal cumulative distribution function and $\boldsymbol{\beta}_j = (\beta_0, ..., \beta_j)$ is a vector of dimension $j + 1$. The first component of the vector x_i is set equal to 1 so that when considering models of the form (10), the intercept is in any submodel. The maximum length of the vector of covariates is $p + 1$. Let $\pi(\boldsymbol{\beta})$, proper or improper, summarize our prior information about $\boldsymbol{\beta}$. Then the posterior density of $\boldsymbol{\beta}$ is given by

$$\pi(\boldsymbol{\beta}|y) = \frac{\pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(x_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(x_i'\boldsymbol{\beta})^{1-y_i})}{\int \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(x_i'\boldsymbol{\beta})^{y_i}(1 - \Phi(x_i'\boldsymbol{\beta})^{1-y_i}) d\boldsymbol{\beta}} \quad ,$$

which is largely intractable.

As shown by [17], the Bayesian probit regression model becomes tractable when a particular set of auxiliary variables is introduced. Based on the data augmentation approach [18], introducing n latent variables $Z_1, ..., Z_n$, where

$$Z_i|\boldsymbol{\beta} \sim N(x_i'\boldsymbol{\beta}, 1).$$

The probit model (10) can be thought of as a regression model with incomplete sampling information by considering that only the sign of z_i is observed. More specifically, define $Y_i = 1$ if $Z_i > 0$ and $Y_i = 0$ otherwise. This allows us to write the probability density of y_i given z_i

$$p(y_i|z_i) = \mathbb{I}(z_i > 0)\mathbb{I}(y_i = 1) + \mathbb{I}(z_i \leq 0)\mathbb{I}(y_i = 0).$$

Expansion of the parameter set from $\{\boldsymbol{\beta}\}$ to $\{\boldsymbol{\beta}, \mathbf{Z}\}$ is the key to achieving a tractable solution for variational approximation.

3.2. Development of Intrinsic Prior for Probit Models

For the sample $\mathbf{z} = (z_1, ..., z_n)'$, the null normal model is

$$M_{\mathbf{1}} : \{N_n(\mathbf{z}|\alpha\mathbf{1}_n, \mathbf{I}_n), \pi(\alpha)\}.$$

For a generic model M_j with $j + 1$ regressors, the alternative model is

$$M_j : \{N_n(\mathbf{z}|\mathbf{X}_j\boldsymbol{\beta}_j, \mathbf{I}_n), \pi(\boldsymbol{\beta}_j)\},$$

where the design matrix $\mathbf{X}_j$ has dimensions $n \times (j + 1)$. Intrinsic prior methodology for the linear model was first developed by [19], and was further developed in [20] by using the methods of [21]. This intrinsic methodology gives us an automatic specification of the priors $\pi(\alpha)$ and $\pi(\boldsymbol{\beta})$, starting with the non-informative priors $\pi^N(\alpha)$ and $\pi^N(\boldsymbol{\beta})$ for α and $\boldsymbol{\beta}$, which are both improper and proportional to 1.

The marginal distributions for the sample $\mathbf{z}$ under the null model, and under the alternative model with intrinsic prior, are formally written as

$$p_1(\mathbf{z}) = \int N_n(\mathbf{z}|\alpha\mathbf{1}_n, \mathbf{I}_n)\pi^N(\alpha)d\alpha,$$
$$p_j(\mathbf{z}) = \int\int N_n(\mathbf{z}|\mathbf{X}_j\boldsymbol{\beta}_j, \mathbf{I}_n)\pi^I(\boldsymbol{\beta}|\alpha)\pi^N(\alpha)d\alpha d\boldsymbol{\beta}. \tag{11}$$

However, these are marginals of the sample $\mathbf{z}$, but our selection procedure requires us to compute the Bayes factor of model M_j versus the reference model $M_\mathbf{1}$ for the sample $\mathbf{y} = (y_1, ..., y_n)$. To solve this problem, reference [2] proposed to transform the marginal $p_j(\mathbf{z})$ into the marginal $p_j(\mathbf{y})$ by using the probit transformations $y_i = 1(z_i > 0), i = 1, ..., n$. These latter marginals are given by

$$p_j(\mathbf{y}) = \int_{A_1\times...\times A_n} p_j(\mathbf{z})d\mathbf{z} \tag{12}$$

where

$$A_i = \begin{cases} (0, \infty) \text{ if } y_i = 1, \\ (-\infty, 0) \text{ if } y_i = 0. \end{cases} \tag{13}$$

4. Variational Inference

4.1. Overview of Variational Methods

Variational methods have their origins in the 18th century with the work of Euler, Lagrange, and others on the calculus of variations (The derivation in this section is standard in the literature on variational approximation and will at times follow the arguments in [22,23]). Variational inference is a body of deterministic techniques for making approximate inference for parameters in complex statistical models. Variational approximations are a much faster alternative to Markov Chain Monte Carlo (MCMC), especially for large models, and are a richer class of methods than the Laplace approximation [6].

Suppose we have a Bayesian model and a prior distribution for the parameters. The model may also have latent variables, here we shall denote the set of all latent variables and parameters by $\boldsymbol{\theta}$. In addition, we denote the set of all observed variables by $\mathbf{X}$. Given a set of n independent, identically distributed data, for which $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_n\}$, our probabilistic model (e.g., probit regression model) specifies the joint distribution $p(\mathbf{X}, \boldsymbol{\theta})$, and our goal is to find an approximation for the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ as well as for the marginal likelihood $p(\mathbf{X})$. For any probability distribution $q(\boldsymbol{\theta})$, we have the following decomposition of the log marginal likelihood

$$\ln p(\mathbf{X}) = \mathcal{L}(q) + \text{KL}(q||p)$$

where we have defined

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \ln \left\{ \frac{p(\mathbf{X}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \tag{14}$$

$$\text{KL}(q||p) = -\int q(\boldsymbol{\theta}) \ln \left\{ \frac{p(\boldsymbol{\theta}|\mathbf{X})}{q(\boldsymbol{\theta})} \right\} d\boldsymbol{\theta} \tag{15}$$

We refer to (14) as the lower bound of the log marginal likelihood with respect to the density q, and (15) is by definition the Kullback–Leibler divergence of the posterior $q(\boldsymbol{\theta}|\mathbf{X})$ from the density q. Based on this decomposition, we can maximize the lower bound $\mathcal{L}(q)$ by optimization with respect to the distribution $q(\boldsymbol{\theta})$, which is equivalent to minimizing the KL divergence. In addition, the lower bound is attained when the KL divergence is zero, which happens when $q(\boldsymbol{\theta})$ equals the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$. It would be hard to find such a density since the true posterior distribution is intractable.

4.2. Factorized Distributions

The essence of the variational inference approach is approximation to the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$ by $q(\boldsymbol{\theta})$ for which the q dependent lower bound $\mathcal{L}(q)$ is more tractable than the original model evidence. In addition, tractability is achieved by restricting q to a more manageable class of distributions, and then maximizing $\mathcal{L}(q)$ over that class.

Suppose we partition elements of $\boldsymbol{\theta}$ into disjoint groups $\{\boldsymbol{\theta}_i\}$ where $i = 1, ..., M$. We then assume that the q density factorizes with respect to this partition, i.e.,

$$q(\boldsymbol{\theta}) = \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i). \tag{16}$$

The product form is the only assumption we made about the distribution. Restriction (16) is also known as *mean-field* approximation and has its root in Physics [24].

For all distributions $q(\boldsymbol{\theta})$ with the form (16), we need to find the distribution for which the lower bound $\mathcal{L}(q)$ is largest. Restriction of q to a subclass of product densities like (16) gives rise to explicit solutions for each product component in terms of the others. This fact, in turn, leads to an iterative scheme for obtaining the solutions. To achieve this, we first substitute (16) into (14) and then separate out the dependence on one of the factors $q_j(\boldsymbol{\theta}_j)$. Denoting $q_j(\boldsymbol{\theta}_j)$ by q_j to keep the notation clear, we obtain

$$\begin{aligned}\mathcal{L}(q) &= \int \prod_{i=1}^{M} q_i \Big\{ \ln p(\mathbf{X}, \boldsymbol{\theta}) - \sum_{i=1}^{M} \ln q_i \Big\} d\boldsymbol{\theta} \\ &= \int q_j \Big\{ \int \ln p(\mathbf{X}, \boldsymbol{\theta}) \prod_{i \neq j} q_i d\boldsymbol{\theta}_i \Big\} d\boldsymbol{\theta}_j - \int q_j \ln q_j d\boldsymbol{\theta}_j + \text{constant} \\ &= \int q_j \ln \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j) d\boldsymbol{\theta}_j - \int q_j \ln q_j d\boldsymbol{\theta}_j + \text{constant}\end{aligned} \tag{17}$$

where $\tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$ is given by

$$\ln \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] + \text{constant}. \tag{18}$$

The notation $\mathbb{E}_{i \neq j}[\cdot]$ denotes an expectation with respect to the q distributions over all variables $\mathbf{z}_i$ for $i \neq j$, so that

$$\mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] = \int \ln p(\mathbf{X}, \boldsymbol{\theta}) \prod_{i \neq j} q_i d\boldsymbol{\theta}_i.$$

Now suppose we keep the $\{q_{i \neq j}\}$ fixed and maximize $\mathcal{L}(q)$ in (17) with respect to all possible forms for the density $q_j(\boldsymbol{\theta}_j)$. By recognizing that (17) is the negative KL divergence between $\tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$ and $q_j(\boldsymbol{\theta}_j)$, we notice that maximizing (17) is equivalent to minimize the KL divergence, and the minimum occurs when $q_j(\boldsymbol{\theta}_j) = \tilde{p}(\mathbf{X}, \boldsymbol{\theta}_j)$. The optimal $q_j^*(\boldsymbol{\theta}_j)$ is then

$$\ln q_j^*(\boldsymbol{\theta}_j) = \mathbb{E}_{i \neq j}[\ln p(\mathbf{X}, \boldsymbol{\theta})] + \text{constant}. \tag{19}$$

The above solution says that the log of the optimal q_j is obtained simply by considering the log of the joint distribution of all parameter, latent and observable variables and then taking the expectation with respect to all the other factors q_i for $i \neq j$. Normalizing the exponential of (19), we have

$$q_j^*(\boldsymbol{\theta}_j) = \frac{\exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i\neq j}[\ln p(\mathbf{X},\boldsymbol{\theta})])d\boldsymbol{\theta}_j}.$$

The set of equations in (19) for $j = 1, ..., M$ are not an explicit solution because the expression on the right hand side of (19) for the optimal q_j^* depends on expectations taken with respect to the other factors q_i for $i \neq j$. We will need to first initialize all of the factors $q_i(\boldsymbol{\theta}_i)$ and then cycle through the factors one by one and replace each in turn with an updated estimate given by the right hand side of (19) evaluated using the current estimates for all of the other factors. Convexity properties can be used to show that convergence to at least local optima is guaranteed [25]. The iterative procedure is described in Algorithm 1.

Algorithm 1 Iterative procedure for obtaining the optimal densities under factorized density restriction (16). The updates are based on the solutions given by (19).

```
1: Initialize $q_2^*(\boldsymbol{\theta}_2), ..., q_M^*(\boldsymbol{\theta}_M)$.
2: Cycle through
       $q_1^*(\boldsymbol{\theta}_1) \leftarrow \frac{\exp(\mathbb{E}_{i\neq 1}[\ln p(\mathbf{X},\boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i\neq 1}[\ln p(\mathbf{X},\boldsymbol{\theta})])d\boldsymbol{\theta}_1}$
       $\vdots$
       $q_M^*(\boldsymbol{\theta}_M) \leftarrow \frac{\exp(\mathbb{E}_{i\neq M}[\ln p(\mathbf{X},\boldsymbol{\theta})])}{\int \exp(\mathbb{E}_{i\neq M}[\ln p(\mathbf{X},\boldsymbol{\theta})])d\boldsymbol{\theta}_M}$
   until the increase in $\mathcal{L}(q)$ is negligible.
```

5. Incorporate Intrinsic Prior with Variational Approximation to Bayesian Probit Models

5.1. Derivation of Intrinsic Prior to Be Used in Variational Inference

Let $\mathbf{X}_l$ be the design matrix of a minimal training sample (mTS) of a normal regression model M_j for the variable $\mathbf{Z} \sim N(\mathbf{X}_j\boldsymbol{\beta}_j, \mathbf{I}_{j+1})$. We have, for the $j+1$-dimensional parameter $\boldsymbol{\beta}_j$,

$$\int N_{j+1}(\mathbf{z}_l|\mathbf{X}_l\boldsymbol{\beta}_j, \mathbf{I}_{j+1})d\boldsymbol{\beta}_j = \begin{cases} |\mathbf{X}_l'\mathbf{X}_l|^{-1/2} & \text{if rank of } \mathbf{X}_l \geq j+1 \\ \infty & \text{otherwise} \end{cases}.$$

Therefore, it follows that the mTS size is $j+1$ [2]. Given that priors for α and $\boldsymbol{\beta}$ are proportional to 1, the intrinsic prior for β conditional on α could be derived. Let $\boldsymbol{\beta}_0$ denote the vector with the first component equal to α and the others equal to zero. Based on Formula (9), we have

$$\begin{aligned}
\pi^I(\boldsymbol{\beta}|\alpha) &= \pi_j^N(\boldsymbol{\beta})\mathbb{E}_{\mathbf{z}_l|\boldsymbol{\beta}}^{M_j}\Big[\frac{p_1(\mathbf{z}_l|\alpha)}{\int p_j(\mathbf{z}_l|\boldsymbol{\beta})\pi_j^N(\boldsymbol{\beta})d\boldsymbol{\beta}}\Big] \\
&= \mathbb{E}_{\mathbf{z}_l|\boldsymbol{\beta}}^{M_j}\Big[\frac{\exp\{-\frac{1}{2}(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta}_0)'(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta}_0)\}}{\int \exp\{-\frac{1}{2}(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta})'(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta})\}d\boldsymbol{\beta}}\Big] \\
&= (2\pi)^{-\frac{(j+1)}{2}}|(\mathbf{X}_l'\mathbf{X}_l)^{-1}|^{-\frac{1}{2}} \times \mathbb{E}_{\mathbf{z}_l|\boldsymbol{\beta}}^{M_j}\Big[\exp\{-\frac{1}{2}(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta}_0)'(\mathbf{z}_l - \mathbf{X}_l\boldsymbol{\beta}_0)\}\Big] \\
&= (2\pi)^{-\frac{(j+1)}{2}}|2(\mathbf{X}_l'\mathbf{X}_l)^{-1}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}[(\boldsymbol{\beta} - \boldsymbol{\beta}_0)'\frac{\mathbf{X}_l'\mathbf{X}_l}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}_0)]\}.
\end{aligned}$$

Therefore,

$$\pi^I(\boldsymbol{\beta}|\alpha) = N_{j+1}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, 2(\mathbf{X}_l'\mathbf{X}_l)^{-1}), \text{ where } \boldsymbol{\beta}_0 = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}_{(j+1)\times 1} .$$

Notice that $\mathbf{X}_l'\mathbf{X}_l$ is unknown because it is a theoretical design matrix corresponding to the training sample $\mathbf{z}_l$. It can be estimated by averaging over all submatrices containing $j+1$ rows of the $n \times (j+1)$ design matrix $\mathbf{X}_j$. This average is $\frac{j+1}{n}\mathbf{X}_j'\mathbf{X}_j$ (See [26] and Appendix A in [2]), and therefore

$$\pi^I(\boldsymbol{\beta}|\alpha) = N_{j+1}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \frac{2n}{j+1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}).$$

Next, based on $\pi^I(\boldsymbol{\beta}|\alpha)$, the intrinsic prior for $\boldsymbol{\beta}$ can be obtained by

$$\pi^I(\boldsymbol{\beta}) = \int \pi^I(\boldsymbol{\beta}|\alpha)\pi^I(\alpha)d\alpha. \tag{20}$$

Since we assume that $\pi^I(\alpha) = \pi^N(\alpha)$ is proportional to one, set $\pi^N(\alpha) = c$ where c is an arbitrary positive constant. Denote $\frac{2n}{j+1}(\mathbf{X}_j'\mathbf{X}_j)^{-1}$ by $\Sigma_{\boldsymbol{\beta}|\alpha}$, we obtain

$$\begin{aligned} \pi^I(\boldsymbol{\beta}) &= \int \pi^I(\boldsymbol{\beta}|\alpha)\pi^I(\alpha)d\alpha \\ &= c \cdot (2\pi)^{-\frac{j+1}{2}}|\Sigma_{\boldsymbol{\beta}|\alpha}|^{-\frac{1}{2}} \int \exp\{-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\}d\alpha \\ &\propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta}\} \times \int \exp\{-\frac{1}{2}[\boldsymbol{\beta}_0'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta}_0 - 2\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta}_0]\}d\alpha \\ &\propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta}\} \times \int \exp\{-\frac{1}{2}(\Sigma_{\boldsymbol{\beta}|\alpha_{(1,1)}}^{-1}\alpha^2 - 2\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha_{(\cdot 1)}}^{-1}\alpha)\}d\alpha \end{aligned} \tag{21}$$

where $\Sigma_{\boldsymbol{\beta}|\alpha_{(1,1)}}^{-1}$ is component of $\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}$ at position row 1 column 1 and $\Sigma_{\boldsymbol{\beta}|\alpha_{(\cdot 1)}}^{-1}$ is the first column of $\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}$. Denote $\Sigma_{\boldsymbol{\beta}|\alpha_{(1,1)}}^{-1}$ by σ_{11} and $\Sigma_{\boldsymbol{\beta}|\alpha_{(\cdot 1)}}^{-1}$ by γ_1, we then obtain

$$\begin{aligned} \pi^I(\boldsymbol{\beta}) &\propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta}\} \times \int \exp\{-\frac{1}{2}\sigma_{11}(\alpha - \frac{\boldsymbol{\beta}'\gamma_1}{\sigma_{11}})^2 + \frac{1}{2}\frac{(\boldsymbol{\beta}'\gamma_1)^2}{\sigma_{11}}\}d\alpha \\ &\propto \exp\{-\frac{1}{2}(\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}'\frac{\gamma_1\gamma_1'}{\sigma_{11}}\boldsymbol{\beta})\} \times \sqrt{2\pi}\sigma_{11}^{-1/2} \\ &\propto \exp\{-\frac{1}{2}\boldsymbol{\beta}'(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1} - \frac{\gamma_1\gamma_1'}{\sigma_{11}})\boldsymbol{\beta}\}. \end{aligned} \tag{22}$$

Therefore, we have derived that

$$\pi^I(\boldsymbol{\beta}) \propto N_{j+1}(\mathbf{0}, (\Sigma_{\boldsymbol{\beta}|\alpha}^{-1} - \frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1}). \tag{23}$$

For model comparison, the specific form of the intrinsic prior may be needed, including the constant factor. Therefore, by following (21) and (22) we have

$$\begin{aligned}\pi^I(\boldsymbol{\beta}) &= c\cdot(2\pi)^{-\frac{j+1}{2}}|\Sigma_{\boldsymbol{\beta}|\alpha}|^{-\frac{1}{2}}(2\pi)^{\frac{j+1}{2}}|(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1}|^{\frac{1}{2}}\sqrt{2\pi}\sigma_{11}^{-1/2}\times N_{j+1}(\mathbf{0},(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1})\\ &= c\cdot|\Sigma_{\boldsymbol{\beta}|\alpha}(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})|^{-\frac{1}{2}}\sqrt{2\pi}\sigma_{11}^{-1/2}\times N_{j+1}(\mathbf{0},(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1})\\ &= c\cdot\sqrt{2\pi}\sigma_{11}^{-1/2}|(\mathbb{I}-\frac{\gamma_1\gamma_1'}{\sigma_{11}}\Sigma_{\boldsymbol{\beta}|\alpha})|^{-\frac{1}{2}}\times N_{j+1}(\mathbf{0},(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1}).\end{aligned} \tag{24}$$

5.2. *Variational Inference for Probit Model with Intrinsic Prior*

5.2.1. Iterative Updates for Factorized Distributions

We have that

$$\begin{aligned}&Z_i|\boldsymbol{\beta}\sim N(x_i'\boldsymbol{\beta},1)\quad\text{and}\\ &p(y_i|z_i)=\mathbb{I}(z_i>0)\mathbb{I}(y_i=1)+\mathbb{I}(z_i\le 0)\mathbb{I}(y_i=0)\end{aligned}$$

in Section 3.1. We have shown in Section 5.1 that

$$\pi^I(\boldsymbol{\beta})\propto N_{j+1}(\mu_{\boldsymbol{\beta}},\Sigma_{\boldsymbol{\beta}}),$$

where $\mu_{\boldsymbol{\beta}}=\mathbf{0}$ and $\Sigma_{\boldsymbol{\beta}}=(\Sigma_{\boldsymbol{\beta}|\alpha}^{-1}-\frac{\gamma_1\gamma_1'}{\sigma_{11}})^{-1}$. Since $\mathbf{y}$ is independent of $\boldsymbol{\beta}$ given $\mathbf{z}$, we have

$$\begin{aligned}p(\mathbf{y},\mathbf{z},\boldsymbol{\beta})&=p(\mathbf{y}|\mathbf{z},\boldsymbol{\beta})p(\mathbf{z}|\boldsymbol{\beta})p(\boldsymbol{\beta})\\ &=p(\mathbf{y}|\mathbf{z})p(\mathbf{z}|\boldsymbol{\beta})p(\boldsymbol{\beta}).\end{aligned} \tag{25}$$

To apply the variational approximation to probit regression model, unobservable variables are considered in two separate groups, coefficient parameter $\boldsymbol{\beta}$ and auxiliary variable $\mathbf{Z}$. To approximate the posterior distribution of $\boldsymbol{\beta}$, consider the product form

$$q(\mathbf{Z},\boldsymbol{\beta})=q_{\mathbf{Z}}(\mathbf{Z})q_{\boldsymbol{\beta}}(\boldsymbol{\beta}).$$

We proceed by first describing the distribution for each factor of the approximation, $q_{\mathbf{Z}}(\mathbf{Z})$ and $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. Then variational approximation is accomplished by iteratively updating the parameters of each factor distribution.

Start with $q_{\mathbf{Z}}(\mathbf{Z})$, when $y_i=1$, we have

$$\log p(\mathbf{y},\mathbf{z},\boldsymbol{\beta})=\log\Big(\prod_i\frac{1}{\sqrt{2\pi}}\exp\{-\frac{(z_i-x_i'\boldsymbol{\beta})^2}{2}\}\times\pi^I(\boldsymbol{\beta})\Big)\qquad\text{where } z_i>0.$$

Now, according to (19) and Algorithm 1, the optimal $q_{\mathbf{Z}}$ is proportional to

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\beta}}[\log p(\mathbf{y},\mathbf{z},\boldsymbol{\beta})]&=-\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}}[\mathbf{z}'\mathbf{z}-2\boldsymbol{\beta}'\mathbf{X}\mathbf{z}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]+\mathbb{E}_{\boldsymbol{\beta}}[\log\pi^I(\boldsymbol{\beta})]\\ &=-\frac{1}{2}\mathbf{z}'\mathbf{z}+\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}]'\mathbf{X}'\mathbf{z}+\cancelto{\text{constant}}{-\frac{1}{2}\mathbb{E}_{\boldsymbol{\beta}}[\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]}+\cancelto{\text{constant}}{\mathbb{E}_{\boldsymbol{\beta}}[\log\pi^I(\boldsymbol{\beta})]}.\end{aligned}$$

So, we have the optimal $q_{\mathbf{Z}}$,

$$
\begin{aligned}
q^*_{\mathbf{Z}}(\mathbf{Z}) &\propto \exp\{-\frac{1}{2}\mathbf{z}'\mathbf{z} + \mathbb{E}_{\beta}[\boldsymbol{\beta}]'\mathbf{X}'\mathbf{z} + \text{constant}\} \\
&\propto \exp\{-\frac{1}{2}(\mathbf{z} - \mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}])'(\mathbf{z} - \mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}])\}.
\end{aligned}
$$

Similar procedure could be used to develop cases when $y_i = 0$. Therefore, we have that the optimal approximation for $q_{\mathbf{Z}}$ is a truncated normal distribution, where

$$
q^*_{\mathbf{Z}}(\mathbf{Z}) = \begin{cases} N_{[0,+\infty)}(\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i, 1) & \text{if } y_i = 1, \\ N_{(-\infty,0]}(\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i, 1) & \text{if } y_i = 0. \end{cases} \tag{26}
$$

Denote $\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]$ by $\mu_{\mathbf{z}}$, the location of distribution $q^*_{\mathbf{Z}}(\mathbf{Z})$. The expectation $\mathbb{E}_{\beta}$ is taken with respect to the density form of $q(\boldsymbol{\beta})$ for which we shall derive now.

For $q_{\beta}(\boldsymbol{\beta})$, given the joint form in (25), we have

$$
\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta}) = -\frac{1}{2}\exp\{(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\} - \frac{1}{2}\exp\{(\boldsymbol{\beta} - \mu_{\beta})'\Sigma_{\beta}^{-1}(\boldsymbol{\beta} - \mu_{\beta})\} + \textbf{constant}.
$$

Taking expectation with respect to $q_{\mathbf{Z}}(\mathbf{z})$, we have

$$
\begin{aligned}
\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta})] = &-\cancelto{\text{constant}}{\frac{1}{2}\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}'\mathbf{Z}]} + \mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]'\mathbf{X}\boldsymbol{\beta} - \frac{1}{2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&- \frac{1}{2}\boldsymbol{\beta}'\Sigma_{\beta}^{-1}\boldsymbol{\beta} + \mu_{\beta}'\Sigma_{\beta}^{-1}\boldsymbol{\beta} + \cancelto{\text{constant}}{\mu_{\beta}'\Sigma_{\beta}^{-1}\mu_{\beta}}.
\end{aligned}
$$

Again, based on (19) and Algorithm 1, the optimal $q_{\beta}(\boldsymbol{\beta})$ is proportional to $\mathbb{E}_{\mathbf{Z}}[\log p(\mathbf{y}, \mathbf{z}, \boldsymbol{\beta})]$,

$$
q^*_{\beta}(\boldsymbol{\beta}) \propto -\frac{1}{2}\boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1})\boldsymbol{\beta} + (\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]'\mathbf{X} + \mu_{\beta}'\Sigma_{\beta}^{-1})\boldsymbol{\beta}.
$$

First notice that any constant terms, including constant factor in the intrinsic prior, were canceled out due to the ratio form of (19). Then by noticing the quadratic form in the above formula we have

$$
q^*_{\beta}(\boldsymbol{\beta}) = N(\mu_{q_{\beta}}, \Sigma_{q_{\beta}}), \tag{27}
$$

where

$$
\begin{aligned}
\Sigma_{q_{\beta}} &= (\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1})^{-1}, \\
\mu_{q_{\beta}} &= (\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1})^{-1}(\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]'\mathbf{X} + \mu_{\beta}'\Sigma_{\beta}^{-1}).
\end{aligned}
$$

Notice that $\mu_{q_{\beta}}$, i.e., $\mathbb{E}_{\beta}[\boldsymbol{\beta}]$, depends on $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]$. In addition, from our previous derivation, we found that the update for $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]$ depends on $\mathbb{E}_{\beta}[\boldsymbol{\beta}]$. Given that the density form of $q_{\mathbf{Z}}$ is truncated normal, we have

$$
\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}_i] = \begin{cases} \mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i + \frac{\phi(-\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i)}{1-\Phi(-\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}])_i} & \text{if } y_i = 1, \\ \mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i - \frac{\phi(-\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}]_i)}{\Phi(-\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}])_i} & \text{if } y_i = 0, \end{cases}
$$

where ϕ is the standard normal density and Φ is the standard normal cumulative density. Denote $\mathbb{E}_{\mathbf{Z}}[\mathbf{Z}]$ by $\mu_{q_{\mathbf{Z}}}$. See properties of truncated normal distribution in Appendix A. Updating procedures for parameters $\mu_{q_{\beta}}$ and $\mu_{q_{\mathbf{Z}}}$ of each factor distribution are summarized in Algorithm 2.

Algorithm 2 Iterative procedure for updating parameters to reach optimal factor densities q^*_{β} and $q^*_{\mathbf{Z}}$ in Bayesian probit regression model. The updates are based on the solutions given by (26) and (27).

```
Initialize $\mu_{q_{\mathbf{Z}}}$.
Cycle through
    $\mu_{q_{\beta}} \leftarrow (\mathbf{X}'\mathbf{X} + \Sigma_{\beta}^{-1})^{-1}(\mu'_{q_{\mathbf{Z}}}\mathbf{X} + \mu'_{\beta}\Sigma_{\beta}^{-1}),$
    $\mu_{q_{\mathbf{Z}}} \leftarrow \mathbf{X}\mu_{q_{\beta}} + \frac{\phi(\mathbf{X}\mu_{q_{\beta}})}{\Phi(\mathbf{X}\mu_{q_{\beta}})^{\mathbf{y}}[\Phi(\mathbf{X}\mu_{q_{\beta}}) - \mathbf{1}]^{\mathbf{1}-\mathbf{y}}} \quad ,$
until the increase in $\mathcal{L}(q)$ is negligible.
```

5.2.2. Evaluation of the Lower Bound $\mathcal{L}(q)$

During the process of optimization of variational approximation densities, the lower bound for the log marginal likelihood need to be evaluated and monitored to determine when the iterative updating process converges. Based on derivations from previous section, we now have the exact form for the variational inference density,

$$q(\boldsymbol{\beta}, \mathbf{Z}) = q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z}).$$

According to (14), we can write down the lower bound $\mathcal{L}(q)$ with respect to $q(\boldsymbol{\beta}, \mathbf{Z})$.

$$\begin{aligned}
\mathcal{L}(q) &= \int q(\boldsymbol{\beta}, \mathbf{Z}) \log\left\{\frac{p(\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z})}{q(\boldsymbol{\beta}, \mathbf{Z})}\right\} d\boldsymbol{\beta} d\mathbf{Z} \\
&= \int q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z}) \log\left\{\frac{p(\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z})}{q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z})}\right\} d\boldsymbol{\beta} d\mathbf{Z} \\
&= \int q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z}) \log\{p(\mathbf{Y}, \boldsymbol{\beta}, \mathbf{Z})\} d\boldsymbol{\beta} d\mathbf{Z} - \int q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z}) \log\{q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z})\} d\boldsymbol{\beta} d\mathbf{Z} \\
&= \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\beta})\}] + \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\pi^{I}(\boldsymbol{\beta})] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{q_{\beta}(\boldsymbol{\beta})\}] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{q_{\mathbf{Z}}(\mathbf{Z})\}].
\end{aligned} \tag{28}$$

As we can see in (28), $\mathcal{L}(q)$ has been divided into four different parts with expectation taken over the variational approximation density $q(\boldsymbol{\beta}, \mathbf{Z}) = q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z})$. We now find the expression of these expectations one by one.

Part 1: $\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{p(\mathbf{Y}, \mathbf{Z}|\boldsymbol{\beta})\}]$

$$\begin{aligned}
&= \log(2\pi)^{-\frac{n}{2}} + \int\int q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{Z})\{-\frac{1}{2}(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta})\} d\boldsymbol{\beta} d\mathbf{z} \\
&= \log(2\pi)^{-\frac{n}{2}} + \int q_{\mathbf{Z}}(\mathbf{Z}) \int q_{\beta}(\boldsymbol{\beta})\{-\frac{1}{2}(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{z}'\mathbf{z})\} d\boldsymbol{\beta} d\mathbf{z}
\end{aligned} \tag{29}$$

Deal with the inner integral first, we have

$$\begin{aligned}
\int q_{\beta}(\boldsymbol{\beta})\{-\frac{1}{2}(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} - 2\mathbf{z}'\mathbf{X}\boldsymbol{\beta} + \mathbf{z}'\mathbf{z})\} d\boldsymbol{\beta} &= -\frac{1}{2}\int q_{\beta}(\boldsymbol{\beta})[\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}] d\boldsymbol{\beta} + \mathbf{z}'\mathbf{X}\mathbb{E}_{\beta}[\boldsymbol{\beta}] - \frac{1}{2}\mathbf{z}'\mathbf{z} \\
&= -\frac{1}{2}\int q_{\beta}(\boldsymbol{\beta})[\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}] d\boldsymbol{\beta} + \mathbf{z}'\mathbf{X}\mu_{q_{\beta}} - \frac{1}{2}\mathbf{z}'\mathbf{z}
\end{aligned} \tag{30}$$

where

$$\begin{aligned}
-\frac{1}{2}\int q_{\beta}(\boldsymbol{\beta})[\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}]d\boldsymbol{\beta} &= -\frac{1}{2}\int q_{\beta}(\boldsymbol{\beta})[(\boldsymbol{\beta}-\mu_{q_{\beta}}+\mu_{q_{\beta}})'\mathbf{X}'\mathbf{X}(\boldsymbol{\beta}-\mu_{q_{\beta}}+\mu_{q_{\beta}})]d\boldsymbol{\beta} \\
&= -\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}\mathbb{E}_{\beta}[(\boldsymbol{\beta}-\mu_{q_{\beta}})(\boldsymbol{\beta}-\mu_{q_{\beta}})']) - \frac{1}{2}\mu'_{q_{\beta}}\mathbf{X}'\mathbf{X}\mu_{q_{\beta}} \\
&= -\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]).
\end{aligned} \tag{31}$$

Substitute (31) into (30), we got

$$\int q_{\beta}(\boldsymbol{\beta})\{-\frac{1}{2}(\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}-2\mathbf{z}'\mathbf{X}\boldsymbol{\beta}+\mathbf{z}'\mathbf{z})\}d\boldsymbol{\beta} = -\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \mathbf{z}'\mathbf{X}\mu_{q_{\beta}} - \frac{1}{2}\mathbf{z}'\mathbf{z}. \tag{32}$$

Substituting (32) back into (29) gives

$$\begin{aligned}
\mathbb{E}_{\beta,\mathbf{Z}}[\log\{p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\beta})\}] &= \log(2\pi)^{-\frac{n}{2}} + \int q_{\mathbf{Z}}(\mathbf{z})\{-\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \mathbf{z}'\mathbf{X}\mu_{q_{\beta}} - \frac{1}{2}\mathbf{z}'\mathbf{z}\}d\mathbf{z} \\
&= \log(2\pi)^{-\frac{n}{2}} - \frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) - \frac{1}{2}\mathbb{E}_{\mathbf{Z}}[\mathbf{z}'\mathbf{z}] + \mu'_{q_{\mathbf{z}}}\mu_{\mathbf{z}} \\
&= \log(2\pi)^{-\frac{n}{2}} - \frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \mu'_{q_{\mathbf{z}}}\mu_{\mathbf{z}} \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}[1+\mu^2_{\mathbf{z}_i} - \mu_{\mathbf{z}_i}\frac{\phi(-\mu_{\mathbf{z}_i})}{\Phi(-\mu_{\mathbf{z}_i})}]^{\mathbb{I}(y_i=0)}[1+\mu^2_{\mathbf{z}_i} + \mu_{\mathbf{z}_i}\frac{\phi(-\mu_{\mathbf{z}_i})}{1-\Phi(-\mu_{\mathbf{z}_i})}]^{\mathbb{I}(y_i=1)} \\
&= \log(2\pi)^{-\frac{n}{2}} - \frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \mu'_{q_{\mathbf{z}}}\mu_{\mathbf{z}} \\
&\quad - \frac{1}{2}\sum_{i=1}^{n}[1+\mu_{q_{\mathbf{z}_i}}\mu_{\mathbf{z}_i}]^{\mathbb{I}(y_i=0)}[1+\mu_{q_{\mathbf{z}_i}}\mu_{\mathbf{z}_i}]^{\mathbb{I}(y_i=1)} \\
&= \log(2\pi)^{-\frac{n}{2}} - \frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \frac{1}{2}\mu'_{q_{\mathbf{z}}}\mu_{\mathbf{z}} - \frac{n}{2}.
\end{aligned} \tag{33}$$

We applied properties of truncated normal distribution in Appendix B to find the expression of the second moment $\mathbb{E}_{\mathbf{Z}}[\mathbf{z}'\mathbf{z}]$.

Part 2: $\mathbb{E}_{\beta,\mathbf{Z}}[\log q_{\mathbf{Z}}(\mathbf{z})]$

$$\begin{aligned}
&= \int\int q_{\beta}(\boldsymbol{\beta})q_{\mathbf{Z}}(\mathbf{z})\log q_{\mathbf{Z}}(\mathbf{z})d\boldsymbol{\beta}d\mathbf{Z} \\
&= \int q_{\mathbf{Z}}(\mathbf{z})\log q_{\mathbf{Z}}(\mathbf{z})d\mathbf{Z} \\
&= -\frac{n}{2}(\log(2\pi)+1) \\
&\quad + \sum_{i=1}^{n}\{[\log(\Phi(-\mu_{\mathbf{z}_i})) + \mu_{\mathbf{z}_i}\frac{\phi(-\mu_{\mathbf{z}_i})}{2\Phi(-\mu_{\mathbf{z}_i})}]^{\mathbb{I}(y_i=0)}[\log(1-\Phi(-\mu_{\mathbf{z}_i})) - \mu_{\mathbf{z}_i}\frac{\phi(-\mu_{\mathbf{z}_i})}{2(1-\Phi(-\mu_{\mathbf{z}_i}))}]^{\mathbb{I}(y_i=1)}\} \\
&= -\frac{n}{2}(\log(2\pi)+1) - \frac{1}{2}\mu'_{\mathbf{z}}\mu_{\mathbf{z}} + \frac{1}{2}\mu'_{q_{\mathbf{z}}}\mu_{\mathbf{z}} + \sum_{i=1}^{n}\{[\log(\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=0)}[\log(1-\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=1)}\}
\end{aligned} \tag{34}$$

Again, see Appendix B for well-known properties of truncated normal distribution. Now subtracting (34) from (33) we got

$$\begin{aligned}
\mathbb{E}_{\beta,\mathbf{Z}}[\log\{p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\beta})\}] - \mathbb{E}_{\beta,\mathbf{Z}}[\log q_{\mathbf{Z}}(\mathbf{z})] &= -\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\beta}}\mu'_{q_{\beta}}+\Sigma_{q_{\beta}}]) + \frac{1}{2}\mu'_{\mathbf{z}}\mu_{\mathbf{z}} + \\
&\quad \sum_{i=1}^{n}\{[\log(\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=0)}[\log(1-\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=1)}\}.
\end{aligned} \tag{35}$$

Based on the exact expression of the intrinsic prior $\pi^I(\boldsymbol{\beta})$, denoting all constant terms by C, we have

Part 3: $\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log p_{\boldsymbol{\beta}}(\boldsymbol{\beta})]$

$$\begin{aligned} &= \int\int q_{\mathbf{Z}}(\mathbf{z}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \log \pi^I(\boldsymbol{\beta}) d\boldsymbol{\beta} d\mathbf{z} \\ &= \log C - \frac{(j+1)}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| - \frac{1}{2}\int q_{\boldsymbol{\beta}}(\boldsymbol{\beta})[\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}]d\boldsymbol{\beta} \end{aligned} \tag{36}$$

To find the expression for the integral, we have

$$\begin{aligned} \int q_{\boldsymbol{\beta}}(\boldsymbol{\beta})[\boldsymbol{\beta}'\Sigma_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}]d\boldsymbol{\beta} &= \int q_{\boldsymbol{\beta}}(\boldsymbol{\beta})(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}} + \mu_{q_{\boldsymbol{\beta}}})'\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}} + \mu_{q_{\boldsymbol{\beta}}})d\boldsymbol{\beta} \\ &= \mathbb{E}[\text{trace}(\Sigma_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}})(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}})')] + \mu_{q_{\boldsymbol{\beta}}}'\Sigma_{\boldsymbol{\beta}}^{-1}\mu_{q_{\boldsymbol{\beta}}} \\ &= \text{trace}(\Sigma_{\boldsymbol{\beta}}^{-1}\Sigma_{q_{\boldsymbol{\beta}}}) + \mu_{q_{\boldsymbol{\beta}}}'\Sigma_{\boldsymbol{\beta}}^{-1}\mu_{q_{\boldsymbol{\beta}}} \end{aligned} \tag{37}$$

Substituting (37) back into (36), we obtained

$$\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log p_{\boldsymbol{\beta}}(\boldsymbol{\beta})] = \log C - \frac{(j+1)}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| - \frac{1}{2}[\text{trace}(\Sigma_{\boldsymbol{\beta}}^{-1}\Sigma_{q_{\boldsymbol{\beta}}}) + \mu_{q_{\boldsymbol{\beta}}}'\Sigma_{\boldsymbol{\beta}}^{-1}\mu_{q_{\boldsymbol{\beta}}}]. \tag{38}$$

Part 4: $\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log q_{\boldsymbol{\beta}}(\boldsymbol{\beta})]$

$$\begin{aligned} &= \int\int q_{\mathbf{Z}}(\mathbf{z}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) \log q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) d\boldsymbol{\beta} \\ &= -\frac{j+1}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{q_{\boldsymbol{\beta}}}| - \frac{1}{2}\int q_{\boldsymbol{\beta}}(\boldsymbol{\beta})(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}})'\Sigma_{q_{\boldsymbol{\beta}}}^{-1}(\boldsymbol{\beta} - \mu_{q_{\boldsymbol{\beta}}})d\boldsymbol{\beta} \\ &= -\frac{j+1}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_{q_{\boldsymbol{\beta}}}| - \frac{1}{2}\text{trace}(\Sigma_{\boldsymbol{\beta}}^{-1}\Sigma_{\boldsymbol{\beta}}) \\ &= -\frac{j+1}{2}(\log(2\pi) + 1) - \frac{1}{2}\log|\Sigma_{q_{\boldsymbol{\beta}}}| \end{aligned} \tag{39}$$

Combining all four parts together, we get

$$\begin{aligned} \mathcal{L}(q) &= \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\beta})\}] + \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\pi^I(\boldsymbol{\beta})] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{q_{\boldsymbol{\beta}}(\boldsymbol{\beta})\}] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{q_{\mathbf{Z}}(\mathbf{Z})\}] \\ &= \underbrace{-\frac{1}{2}\text{trace}(\mathbf{X}'\mathbf{X}[\mu_{q_{\boldsymbol{\beta}}}\mu_{q_{\boldsymbol{\beta}}}' + \Sigma_{q_{\boldsymbol{\beta}}}]) + \frac{1}{2}\mu_{\mathbf{z}}'\mu_{\mathbf{z}} + \sum_{i=1}^{n}\{[\log(\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=0)}[\log(1-\Phi(-\mu_{\mathbf{z}_i}))]^{\mathbb{I}(y_i=1)}\}}_{\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{p(\mathbf{Y},\mathbf{Z}|\boldsymbol{\beta})\}] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log\{q_{\mathbf{Z}}(\mathbf{Z})\}]} \\ &\quad \underbrace{+\log C - \frac{1}{2}\log|\Sigma_{\boldsymbol{\beta}}| - \frac{1}{2}[\text{trace}(\Sigma_{\boldsymbol{\beta}}^{-1}\Sigma_{q_{\boldsymbol{\beta}}}) + \mu_{q_{\boldsymbol{\beta}}}'\Sigma_{\boldsymbol{\beta}}^{-1}\mu_{q_{\boldsymbol{\beta}}}] + \frac{j+1}{2} + \frac{1}{2}\log|\Sigma_{q_{\boldsymbol{\beta}}}|}_{\mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log p_{\boldsymbol{\beta}}(\boldsymbol{\beta})] - \mathbb{E}_{\boldsymbol{\beta},\mathbf{Z}}[\log q_{\boldsymbol{\beta}}(\boldsymbol{\beta})]}. \end{aligned} \tag{40}$$

5.3. Model Comparison Based on Variational Approximation

Suppose we want to compare two models, M_1 and M_0, where M_0 is the simpler model. An intuitive thought on comparing two models by variational approximation methods is just to compare the lower bounds $\mathcal{L}(q_1)$ and $\mathcal{L}(q_0)$. However, we should note that by comparing the lower bounds, we are assuming that the KL divergences in the two approximations are the same, so that we can use just these lower bounds as guide. Unfortunately, it is not easy to measure how tight in theory any particular bound can be, if this can be accomplished we could then more accurately estimate the log marginal likelihood from the beginning. As clarified in [27], when comparing two exact log marginal likelihood, we have

$$\begin{aligned}\log p_1(\mathbf{X}) - \log p_0(\mathbf{X}) &= [\mathcal{L}(q_1) + KL(q_1 \parallel p_1)] - [\mathcal{L}(q_0) - KL(q_0 \parallel p_0)] \quad (41)\\ &= \mathcal{L}(q_1) - \mathcal{L}(q_0) + [KL(q_1 \parallel p_1) - KL(q_0 \parallel p_0)] \quad (42)\\ &\neq \mathcal{L}(q_1) - \mathcal{L}(q_0). \quad (43)\end{aligned}$$

The difference in log marginal likelihood, $\log p_1(\mathbf{X}) - \log p_0(\mathbf{X})$, is the quantity we wish to estimate. However, if we base this on the lower bounds difference, we are basing our model comparison on (43) rather than (42). Therefore, there exists a systematic bias towards simpler model when comparing models if $KL(q_1 \parallel p_1) - KL(q_0 \parallel p_0)$ is not zero.

Realizing that we have a variational approximation for the posterior distribution of $\boldsymbol{\beta}$, we propose the following method to estimate $p(\mathbf{X})$ based on our variational approximation $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ (27). First, writing the marginal likelihood as

$$p(\mathbf{x}) = \int \Big[\frac{p(\mathbf{x}|\boldsymbol{\beta})\pi^I(\boldsymbol{\beta})}{q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}\Big] q_{\boldsymbol{\beta}}(\boldsymbol{\beta}) d\boldsymbol{\beta},$$

we can interpret it as the conditional expectation

$$p(\mathbf{x}) = \mathbb{E}\Big[\frac{p(\mathbf{x}|\boldsymbol{\beta})\pi^I(\boldsymbol{\beta})}{q_{\boldsymbol{\beta}}(\boldsymbol{\beta})}\Big]$$

with respect to $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. Next, draw samples $\boldsymbol{\beta}^{(1)}, ..., \boldsymbol{\beta}^{(n)}$ from $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ and obtain the estimated marginal likelihood

$$\widehat{p_{\mathbf{X}}(\mathbf{x})} = \frac{1}{n}\sum_{i=1}^{n} \frac{p(\mathbf{x}|\boldsymbol{\beta}^{(i)})\pi^I(\boldsymbol{\beta}^{(i)})}{q_{\boldsymbol{\beta}}(\boldsymbol{\beta}^{(i)})}.$$

Please note that this method proposed is equivalent to importance sampling with importance function being $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$, for which we know the exact form and the generation of the random $\boldsymbol{\beta}^{(i)}$ is easy and inexpensive.

6. Modeling Probability of Default Using Lending Club Data

6.1. Introduction

LendingClub (https://www.lendingclub.com/) is the world's largest peer-to-peer lending platform. LendingClub enables borrowers to create unsecured personal loans between \$1000 and \$40,000. The standard loan period is three or five years. Investors can search and browse the loan listings on LendingClub website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from interest. LendingClub makes money by charging borrowers an origination fee and investors a service fee. To attract lenders, LendingClub publishes most of the information available in borrowers' credit reports as well as information reported by borrowers for almost every loan issued through its website.

6.2. Modeling Probability of Default—Target Variable and Predictive Features

Publicly available LendingClub data, from 2007 June to 2018 Q4, has a total of 2,260,668 issued loans. Each loan has a status, either Paid-off, Charged-off, or Ongoing. We only adopted loans with an end status, i.e., either paid-off or charged-off. In addition, that loan status is the target variable. We then selected following loan features as our predictive covariates.

- Loan term in months (either 36 or 60)
- FICO
- Issued loan amount
- DTI (Debt to income ratio, i.e., customer's total debt divided by income)
- Number of credit lines opened in past 24 months
- Employment length in years
- Annual income
- Home ownership type (own, mortgage, of rent)

We took a sample from the original data set that has customer yearly income between $15,000 and $60,000 and end up with a data set of 520,947 rows.

6.3. Addressing Uncertainty of Estimated Probit Model Using Variational Inference with Intrinsic Prior

Using the process developed in Section 5, we can update the intrinsic prior for parameters (see Figure 1) of the probit model using variational inference, and get the posterior distribution for the estimated parameters. Based on the derived parameter distributions, questions of interest may be explored with model uncertainty being considered.

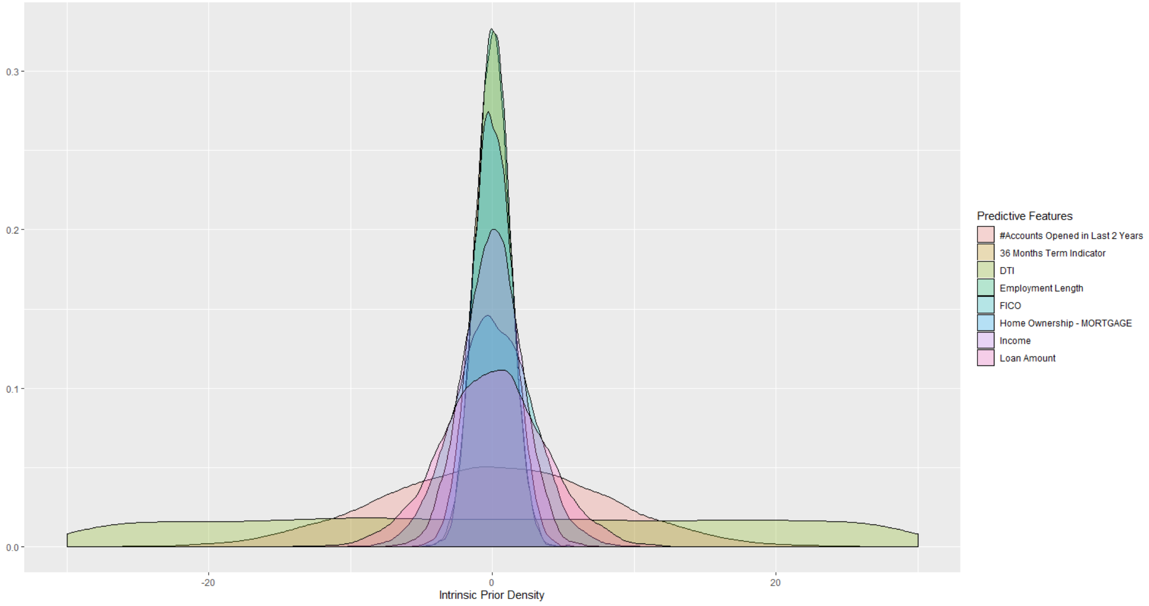

Figure 1. Intrinsic Prior.

Investors will be interested in understanding how each loan feature affect the probability of default, given a certain loan term, either 36 or 60. To answer this question, we samples 6000 cases from the original data set and draw from derived posterior distribution 100 times. We end up with 6000 × 100 calculated probability of default, where each one of the 6000 samples yield 100 different probit estimates based on 100 different posterior draws. We summarize some of our findings in Figure 2, where color red representing 36 months loans and green representing 60 months loans.

- In general, 60 months loans have higher risk of default.
- Given loan term months, there is a clear trend showing that high FICO means lower risk.
- Given loan term months, there is a trend showing that high DTI indicating higher risk.
- Given loan term months, there is a trend showing that more credit lines opened in past 24 months indicating higher risk.

- There is no clear pattern regarding income. This is probably because we only included customers with income between $15,000 and $60,000 in our training data, which may not representing the true income level of the whole population.

Model uncertainty could also be measured through credible intervals. Again, with the derived posterior distribution, the credible interval is just the range containing a particular percentage of estimated effect/parameter values. For instance, the 95% credible interval of the estimated parameter value of FICO is simply the central portion of the posterior distribution that contains 95% of the estimated values. Contrary to the frequentist confidence intervals, Bayesian credible interval is much more straightforward to interpret. Using the Bayesian framework created in this article, from Figure 3, we can simply state that given the observed data, the estimated effect of DTI on default has 89% probability of falling within $[8.300, 8.875]$. Instead of the conventional 95%, we used 89% following suggestions in [28,29], which is just as arbitrary as any of the conventions.

One of the main advantages of using variational inference over MCMC is that variational inference is much faster. Comparisons were made between the two approximation frameworks on a 64-bit Windows 10 laptop, with 32.0 GB RAM. Using the data set introduced in Section 6.2, we have that

- with a conjugate prior and following the Gibbs sampling scheme proposed by [17], it took 89.86 s to finish 100 simulations for the Gibbs sampler;
- following our method proposed in Section 5.2, it took 58.38 s to get the approximated posterior distribution and sampling 10,000 times from that posterior.

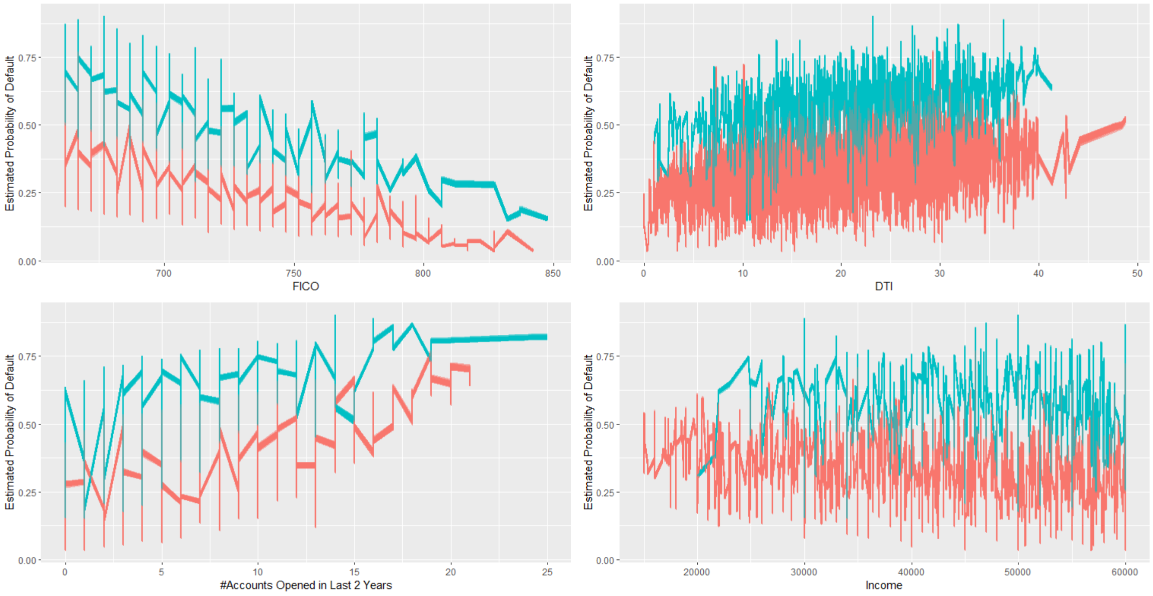

Figure 2. Effect of term months and other covariates on probability of default

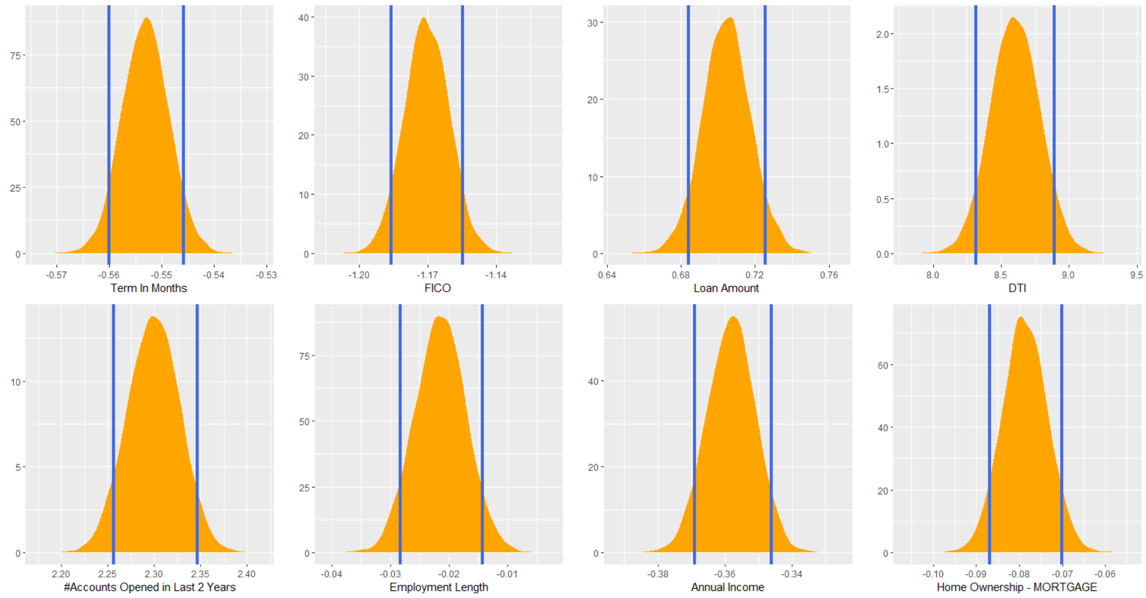

Figure 3. Credible intervals for estimated coefficients

6.4. Model Comparison

Following the procedure proposed in Section 5.3, we compare the following series of nested models. From the data set introduced in Section 6.2, 2000 records were sampled to estimate the likelihood $p(\mathbf{x}|\boldsymbol{\beta}^{(i)})$. Where $\boldsymbol{\beta}^{(i)}$ is one of the 2500 draws sampled directly from the approximated posterior distribution $q_{\beta}(\boldsymbol{\beta})$, which serves as the importance function used to estimate the marginal likelihood $p(\mathbf{x})$.

- M_2: ***FICO*** + ***Term 36 Indicator***
- M_3: ***FICO*** + ***Term 36 Indicator*** + ***Loan Amount***
- M_4: ***FICO*** + ***Term 36 Indicator*** + ***Loan Amount*** + ***Annual Income***
- M_5: ***FICO*** + ***Term 36 Indicator*** + ***Loan Amount*** + ***Annual Income*** + ***Mortgage Indicator***

Estimated log marginal likelihood for each model is plotted in Figure 4. We can see that the model evidence has increased by adding predictive features ***Loan Amount*** and ***Annual Income*** sequentially. However, if we further adding home ownership information, i.e., ***Mortgage Indicator*** as a predictive feature, the model evidence decreased. We have the Bayes factor

$$BF_{45} = \frac{p(\mathbf{x}|M_4)}{p(\mathbf{x}|M_5)} = e^{-1014.78-(-1016.42)} = 5.16,$$

which suggests a substantial evidence for model M_4, indicating home ownership information may be irrelevant in predicting probability of default given that all the other predictive features are relevant.

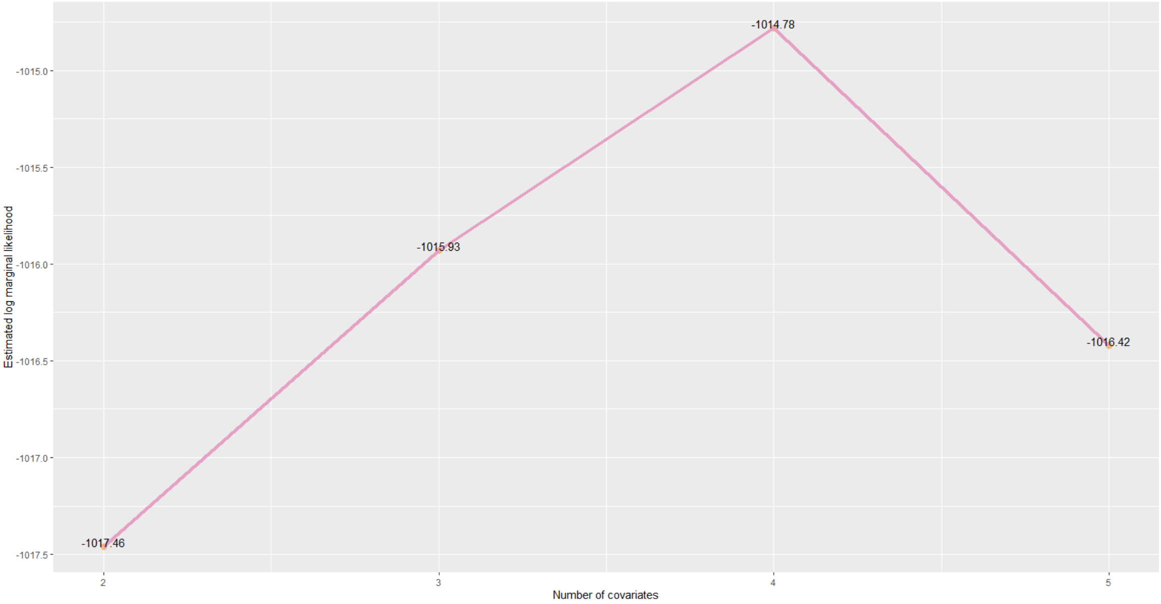

Figure 4. Log marginal likelihood comparison

7. Further Work

The authors thank the reviewers for pointing out that mean-field variational Bayes underestimates the posterior variance. This could be an interesting topic for our future research. We plan to study the *linear response variational Bayes* (LRVB) method proposed in [30] to see if it can be applied on the framework we proposed in this article. To see if we can get the approximated posterior variance close enough to the true variance using our proposed method, comparisons should be made between normal conjugate prior with the MCMC procedure, normal conjugate prior with LRVB, and intrinsic prior with LRVB.

Author Contributions: Methodology, A.L., L.P. and K.W.; software, A.L.; writing–original draft preparation, A.L., L.P. and K.W.; writing–review and editing, A.L. and L.P.; visualization, A.L. All authors have read and agreed to the published version of the manuscript.

Funding: The work of L.R.Pericchi was partially funded by NIH grants U54CA096300, P20GM103475 and R25MD010399.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Density Function

Suppose $X \sim N(\mu, \sigma^2)$ has a normal distribution and lies within the interval $X \in (a, b), -\infty \leq a < b \leq \infty$. Then X conditional on $a < X < b$ has a truncated normal distribution. Its probability density function, f, for $a \leq X < b$, is given by

$$f(x|\mu, \sigma, a, b) = \frac{\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$$

and by $f = 0$ otherwise. Here

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}\xi^2)$$

is the probability density function of the standard normal distribution and $\Phi(\cdot)$ is its cumulative distribution function. If $b = \infty$, then $\Phi(\frac{b-\mu}{\sigma}) = 1$, and similarly, if $a = -\infty$, then $\Phi(\frac{a-\mu}{\sigma}) = 0$. And the cumulative density for the truncated normal distribution is

$$F(x|\mu, \sigma, a, b) = \frac{\Phi(\xi) - \Phi(\alpha)}{Z},$$

where $\xi = \frac{x-\mu}{\sigma}$ and $Z = \Phi(\beta) - \Phi(\alpha)$.

Appendix B. Moments and Entropy

Let $\alpha = \frac{a-\mu}{\sigma}$ and $\beta = \frac{b-\mu}{\sigma}$. For two-sided truncation:

$$\begin{aligned}\mathbb{E}(X|a < X < b) &= \mu + \sigma\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)},\\ Var(X|a < X < b) &= \sigma^2\Big[1 + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{\Phi(\beta) - \Phi(\alpha)} - \Big(\frac{\phi(\alpha) - \phi(\beta)}{\Phi(\beta) - \Phi(\alpha)}\Big)^2\Big].\end{aligned}$$

For one sided truncation (upper tail):

$$\begin{aligned}\mathbb{E}(X|X > a) &= \mu + \sigma\lambda(\alpha)\\ Var(X|X > a) &= \sigma^2[1 - \delta(\alpha)],\end{aligned}$$

where $\alpha = \frac{a-\mu}{\sigma}, \lambda(\alpha) = \frac{\phi(\alpha)}{1-\Phi(\alpha)}$ and $\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$.

For one sided truncation (lower tail):

$$\begin{aligned}\mathbb{E}(X|X < b) &= \mu - \sigma\frac{\phi(\beta)}{\Phi(\beta)}\\ Var(X|X < b) &= \sigma^2\Big[1 - \beta\frac{\phi(\beta)}{\Phi(\beta)} - \Big(\frac{\phi(\beta)}{\Phi(\beta)}\Big)^2\Big].\end{aligned}$$

More generally, the moment generating function for truncated normal distribution is

$$e^{\mu t + \sigma^2 t^2/2} \cdot \Big[\frac{\Phi(\beta - \sigma t) - \Phi(\alpha - \sigma t)}{\Phi(\beta) - \Phi(\alpha)}\Big].$$

For a density $f(x)$ defined over a continuous variable, the *entropy* is given by

$$H[x] = -\int f(x)\log f(x)dx.$$

And the entropy for a truncated normal density is

$$\log(\sqrt{2\pi e}\sigma Z) + \frac{\alpha\phi(\alpha) - \beta\phi(\beta)}{2Z}.$$

References

1. Salmeron, D.; Cano, J.A.; Robert, C.P. Objective Bayesian hypothesis testing in binomial regression models with integral prior distributions. *Stat. Sin.* **2015**, *25*, 1009–1023. [CrossRef]
2. Leon-Novelo, L.; Moreno, E.; Casella, G. Objective Bayes model selection in probit models. *Stat. Med.* **2012**, *31*, 353–365. [CrossRef] [PubMed]
3. Jaakkola, T.S.; Jordan, M.I. Bayesian parameter estimation via variational methods. *Stat. Comput.* **2000**, *10*, 25–37. [CrossRef]

4. Girolami, M.; Rogers, S. Variational Bayesian multinomial probit regression with Gaussian process priors. *Neural Comput.* **2006**, *18*, 1790–1817. [CrossRef]
5. Consonni, G.; Marin, J.M. Mean-field variational approximate Bayesian inference for latent variable models. *Comput. Stat. Data Anal.* **2007**, *52*, 790–798. [CrossRef]
6. Ormerod, J.T.; Wand, M.P. Explaining variational approximations. *Am. Stat.* **2010**, *64*, 140–153. [CrossRef]
7. Grimmer, J. An introduction to Bayesian inference via variational approximations. *Political Anal.* **2010**, *19*, 32–47. [CrossRef]
8. Blei, D.M.; Kucukelbir, A.; McAuliffe, J.D. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.* **2017**, *112*, 859–877. [CrossRef]
9. Pérez, M.E.; Pericchi, L.R.; Ramírez, I.C. The Scaled Beta2 distribution as a robust prior for scales. *Bayesian Anal.* **2017**, *12*, 615–637. [CrossRef]
10. Mulder, J.; Pericchi, L.R. The matrix-*F* prior for estimating and testing covariance matrices. *Bayesian Anal.* **2018**, *13*, 1193–1214. [CrossRef]
11. Berger, J.O.; Pericchi, L.R. Objective Bayesian Methods for Model Selection: Introduction and Comparison. In *Model Selection*; Institute of Mathematical Statistics: Beachwood, OH, USA, 2001; pp. 135–207.
12. Pericchi, L.R. Model selection and hypothesis testing based on objective probabilities and Bayes factors. *Handb. Stat.* **2005**, *25*, 115–149.
13. Scott, J.G.; Berger, J.O. Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Stat.* **2010**, *38*, 2587–2619. [CrossRef]
14. Jeffreys, H. *The Theory of Probability*; OUP: Oxford, UK, 1961.
15. Berger, J.O.; Pericchi, L.R. The intrinsic Bayes factor for model selection and prediction. *J. Am. Stat. Assoc.* **1996**, *91*, 109–122. [CrossRef]
16. Leamer, E.E. *Specification Searches: Ad Hoc Inference with Nonexperimental Data*; Wiley: New York, NY, USA, 1978; Volume 53.
17. Albert, J.H.; Chib, S. Bayesian analysis of binary and polychotomous response data. *J. Am. Stat. Assoc.* **1993**, *88*, 669–679. [CrossRef]
18. Tanner, M.A.; Wong, W.H. The calculation of posterior distributions by data augmentation. *J. Am. Stat. Assoc.* **1987**, *82*, 528–540. [CrossRef]
19. Berger, J.O.; Pericchi, L.R. The intrinsic Bayes factor for linear models. *Bayesian Stat.* **1996**, *5*, 25–44.
20. Casella, G.; Moreno, E. Objective Bayesian variable selection. *J. Am. Stat. Assoc.* **2006**, *101*, 157–167. [CrossRef]
21. Moreno, E.; Bertolino, F.; Racugno, W. An intrinsic limiting procedure for model selection and hypotheses testing. *J. Am. Stat. Assoc.* **1998**, *93*, 1451–1460. [CrossRef]
22. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: Berlin/Heidelberg, Germany, 2006.
23. Jordan, M.I.; Ghahramani, Z.; Jaakkola, T.S.; Saul, L.K. An introduction to variational methods for graphical models. *Mach. Learn.* **1999**, *37*, 183–233. [CrossRef]
24. Parisi, G.; Shankar, R. Statistical field theory. *Phys. Today* **1988**, *41*, 110. [CrossRef]
25. Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: Cambridge, UK, 2004.
26. Berger, J.; Pericchi, L. Training samples in objective Bayesian model selection. *Ann. Stat.* **2004**, *32*, 841–869. [CrossRef]
27. Beal, M.J. *Variational Algorithms for Approximate Bayesian Inference*; University College London: London, UK, 2003.
28. Kruschke, J. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan*; Academic Press: Cambridge, MA, USA, 2014.
29. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018.
30. Giordano, R.J.; Broderick, T.; Jordan, M.I. Linear response methods for accurate covariance estimates from mean field variational Bayes. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, USA, 7–12 December 2015; pp. 1441–1449.