Combining Gaussian Variational Approximation with Laplace Approximation – Simulation Study for Logistic Regression

Jeremy Austin

[Report from an individualized study course at Michigan State University with Dr. Maiti and Professor Bhattacharya]

Introduction and Summary of Recent Findings

The need to approximate a posterior distribution is an ever-present problem in statistics and in practices. Posterior distributions represent the current state of knowledge about uncertain quantities but are often difficult to express. Various methods exist for attempting inference using posterior distributions, one of the most common being varieties of Markov Chain Monte Carlo (MCMC) methods, which can give true samples from a distribution. Unfortunately, MCMC methods tend to struggle when the number of uncertain quantities is very large. It is also difficult to measure whether a chain has stabilized or how many samples is good enough to represent a distribution. Rather than obtaining samples, it is sometimes better to approximate the posterior distribution using a well-known family, like gaussians. Variational approximations are an example of one way to obtain a surrogate to a posterior distribution. The general idea is to minimize a simple divergence measure between the approximation and posterior that can be simplified to not rely on the pesky normalizing constant. The variational approximator can be restricted to follow families like gaussians as well. These types of methods can be contrasted with established methods, like the Laplace approximation, which focuses on matching a gaussian using the gradient and hessian of the unnormalized log posterior.

The Gaussian variational approximation (GVA) satisfies the following condition:

$$\hat{p} = \text{argmin}_{q \in Q}\{D_{KL}(q||p)\}$$

Where Q is the space of all multivariate Gaussians, p is the true posterior, and $\hat{p}$ is the KL-optimal approximation. $D_{KL}(q||p)$ is called the Kullback-Leibler divergence between the approximations and the true posterior:

$$D_{KL}(q||p) = E_q\left[ln\left(\frac{q(\theta)}{p(\theta|D)}\right)\right]$$

This divergence is not symmetric with respect to q and p, and only penalizes q when it fails to put low probability where p has low probability, thus it is not immediately clear that it could produce good posterior approximations. However, over the past few years, some interesting discoveries were made regarding the quality of GVA. In [1], Katsevich and Rigollete found bounds on the total variation distance, mean l2 norm loss, and covariance operator norm loss, between a true posterior satisfying certain conditions and the GVA. In [2], it was found that GVA beats the classical Laplace approximation

(LA) in terms of the posterior mean estimate. LA and GVA also have similar performance when estimating the posterior covariance. Both LA and GVA share some asymptotic guarantees, with the Bernstein Von-Mises theorem justifying LA and Blei and Wang's analogue in [3] justifying GVA.

Despite these incredible findings, the issue of estimating all d(d-1)/2 variance components has a major impact on the applicability of this model. With LA, estimating the covariance components only involves computing the hessian of the unnormalized log posterior and evaluating it at the parameter that maximizes the log posterior (MAP). This means that the estimation part of LA is only of order d instead of d^2 as it is with GVA.

A classical implementation of variational approximation is to choose a factorizable surrogate. This type of surrogate is usually called a mean-field variational approximation and can be specialized to a gaussian case. If a gaussian distribution is factorizable, it means that the off-diagonal components of the covariance matrix are zero. This means if the KL-divergence is used to optimize for a gaussian distribution with these restrictions (MFGVA), it would also be of estimation order d. This sounds great, but unfortunately the KL-divergence fails spectacularly if the true posterior has dependence structures. Recall that the KL-divergence only penalizes q when it fails to put low probability in regions where p has low probability. To see what effect this has, consider the case that q is a factorizable gaussian approximator, and p is an unstructured gaussian.

$$\begin{aligned} D_{KL}(q||p) &= E_q\left[ln\left(\frac{q(\mathrm{x}|\mathrm{m},\mathrm{S})}{p(\mathrm{x}|\mu,\Sigma)}\right)\right] \\ &= \frac{1}{2}(log(|\Sigma|) - log(|S|) + tr(\Sigma^{-1}S) + (m-\mu)'\Sigma^{-1}(m-\mu) - d) \end{aligned}$$

Σ^{-1} is positive-semidefinite, so an optimal m must be μ. Looking at the terms depending on S, differentiating, and setting it to 0, we get:

$$\begin{aligned} \nabla_S D_{KL}(q||p)\} &= \frac{1}{2}\left(\nabla_S tr(\Sigma^{-1}S) - \nabla_S log(|S|)\right) \\ &= \frac{1}{2}\left((\Sigma^{-1} \odot I) - S^{-1}\right) \\ &= 0 \\ \Rightarrow S^{-1} &= (\Sigma^{-1} \odot I) \end{aligned}$$

Where $\odot$ is the element-wise product, and is produced from the fact that S is restricted to only be diagonal. The diagonals of the true precision matrix, Σ^{-1}, turn out to be the inverses of the variances for the full conditional distributions. This means that the mean-field approximation in this case would be a gaussian having the correct mean, but potentially devastatingly small estimated marginal variances, as the full conditional variances can be much smaller than the true marginal variances when Σ has non-zero off diagonals. Even if MFGVA can produce good posterior mean estimates, the covariance estimates are very problematic in general.

The focus of the remainder of this paper is on a proposed alternative to LA, GVA, and MFGVA.

Proposed Model – Structured Gaussian Variational Approximation

The proposed model attempts to leverage both the superior posterior mean estimate of GVA, while maintaining the easy estimation of useful variance components that comes with LA, especially for generalized linear models where the hessian has a simple form. Following the same notation as before, the proposed model, $\hat{p}$, is as follows:

$$\hat{p} = \text{argmin}_{q \in Q}\{D_{KL}(q||p)\}$$

Where Q is the space of multivariate gaussians having the restriction that their precision matrix is proportional to the hessian of the negative unnormalized log posterior at the maximum a posteriori estimate, $\hat{\theta}$, with scale c^{-2} that is part of the optimization.

$$q \in Q$$
$$\Rightarrow \Sigma_q^{-1} = c^{-2}\nabla_\theta^2 V(\hat{\theta})$$

The motivation behind this restriction is from [1] and [2], where the LA covariance matrix is found to have similar operator norm loss order compared to the purely unstructured GVA matrix, with respect to the number of observations and coefficients. While perhaps not improving the loss order, the hope is that training the scale factor improves the overall estimation when the log posterior has skew or higher order curvatures by some constant factor. We found empirically, for the case of logistic regression, the scale factor is estimated to be greater than one and has a slightly improved operator norm loss to the true posterior covariance matrix as compared to the classical Laplace approximation. The next section details these simulations, and this model will hence be referred to as Structured Gaussian Variational Approximation (SGVA).

Simulation Results for Logistic Regression

For all the following results, some assumptions on the data and true coefficients are as follows:

$$X_{i,j=0} = 1$$
$$X_{i,j\neq 0} \sim N(0,1)$$
$$\beta_j \sim N(0,100^2)$$
$$y_i|X,\beta \sim Bern(S(x_i'\beta))$$

$$S(x_i'\beta) = \frac{1}{1 + e^{-x_i'\beta}}$$

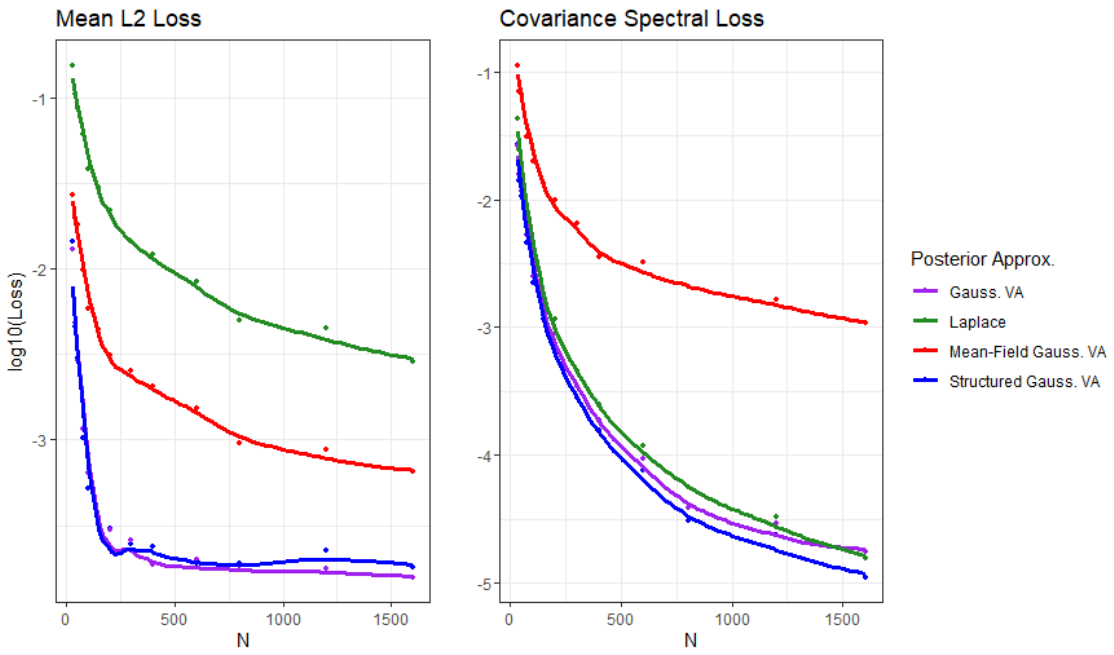
The prior distributions for β_j are intentionally chosen to reflect a lack of knowledge a priori about the location of the true coefficients. Also, the first column of the data matrix is set to 1 to capture an intercept effect. Correlation among the columns of X induces correlation on the posterior distribution for the β_j's; to keep is simple, these columns are assumed to be independent. The logistic regression posterior distribution can be summarized as:

$$p(\beta|D) \propto \prod_{i=1}^{n} [S(x_i'\beta)^{y_i}\left(1 - S(x_i'\beta)\right)^{1-y_i}]\, p(\beta)$$

This model is useful for modeling dependent binary random variables and has a wide range of applications. Many of the closed form solutions of the posterior that are available for linear regression are not available for logistic regression, thus it is important to be able to approximate the posterior using an interpretable distribution like a gaussian.

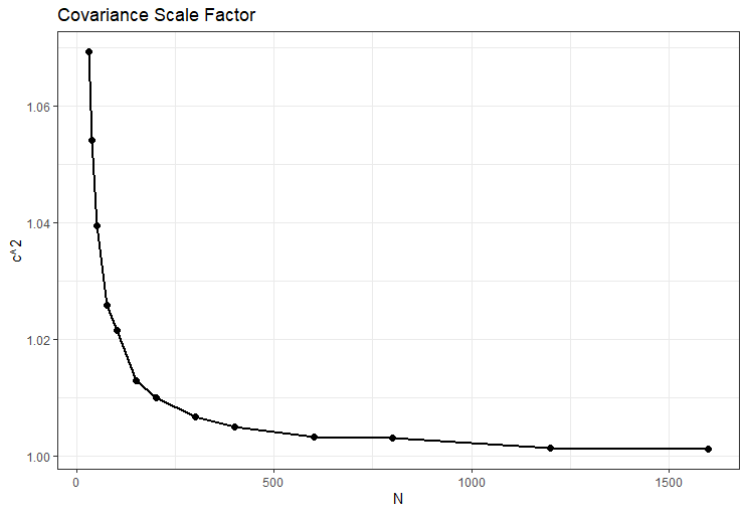
To compare the quality of the posterior approximations introduced in the first section, we first set the number of columns in X to 2, creating just an intercept and slope coefficient to estimate. For each of the approximation methods, the mean of the corresponding surrogate gaussian is compared to an estimate of the true posterior mean. The estimate of the true posterior mean is provided by averaging many MCMC samples provided by STAN. The distances for the means are simply the L2 norm of the difference vectors. The covariance matrices of the surrogate gaussians are also compared to an estimate of the true posterior covariance, once again provided by MCMC. The distances for the covariance matrices are the square root of the largest eigenvalue of the difference matrices. By using these two distance measures, we at least obtain a good look at how well each approximation is doing in terms of two very important moments.

The next image summarizes the mean and covariance estimate qualities. For each sample size, 35 replications of the experiment were done, and the resulting losses were averaged.



The mean estimates for GVA and SGVA are the best, followed by MFGVA, then LA. For the covariance matrices, LA, SGVA, and GVA perform very similarly with SGVA appearing to slightly outperform the rest by some factor. Unsurprisingly, MFGVA does not approximate the covariance matrix well because the off diagonals are 0. The similarity between SGVA and GVA gives some reassurance that some of the recent research regarding GVA quality might also be applicable to SGVA.

The SGVA scale factor also seems to be training to be larger than 1, which is a good sign that it is capturing the width of the posterior distribution a little better than LA, especially for situations where there is little data.



Future Direction

While the results look promising for SGVA, there are no theoretical guarantees for the approximation as there is with GVA and LA. There is a lot of groundwork laid out by researchers like Katsevich, Rigollete, Blei, and Wang that may make it possible to apply their findings to our proposed model. The obvious future direction for this study would be to establish similar theoretical guarantees for this method and rigorously prove its advantages over LA and GVA, while also testing its performance on different types of models and settings.

References

[1] Katsevich and Rigollet. On the Approximation Accuracy of Gaussian Variational Inference. arXiv:2301.02168, 2023.

[2] Katsevich. The Laplace approximation accuracy in high dimensions: a refined analysis and new skew adjustment. arXiv:2306.07262, 2024

[3] Blei and Wang. Frequentist Consistency of Variational Bayes. arXiv:1705.03439v3, 2021