## Honour Presentation Draft

* **Introduction: Slide 1**

Hi everyone, this is Yuhao from the school of mathematic and statistics. My supervisors are Associate Professor: John Ormerod and Doctor Mohammad Davoudabadi. My honor project’s title is A novel algorithm for the Bayesian Lasso: A local approximation adjustment approach, which is basically about designing a more efficient and accurate algorithm for the Bayesian Lasso problem.

* **Motivation: Slide 2**
  + **Here are some of the motivations for this project,**

The Lasso problem is renowned for simultaneously coefficient estimation and regularization, for generating sparse models, but the ordinary lasso is not suitable for obtaining uncertainty quantification. Therefore, a Bayesian lasso model is proposed, which incorporates an exponential prior and the log likelihood.

* + On the other hand, obtaining the posterior of the Bayesian Lasso model involves using Monte Carlo Markov Chain methods that can be computationally demanding.
  + Meanwhile, the variational approximation becomes a faster alternative to Monte Carlo method.
  + Nevertheless, variational approximation does not guarantee finding the exact target density. It sometimes can only find a density close to the target which means the approximation accuracy might be a significant concern.
  + To address the slow speed and imprecise issue of obtaining posterior distribution of the Bayesian Lasso, new methods should be explored.
* **Overview: Slide 3**
  + This presentation is organized as follows. The first and second section illustrates the motivation and background of the Lasso problem and the transition to the Bayesian Lasso Problem. We present our main methodology based on the variational approximation in the third and fourth sections, followed by a comprehensive experiment for testing the effectiveness of the algorithm as well as some discussion, limitations, and future work.
* **Lasso: Formulation (Slide 4)**
  + **Now I’d like to start by introducing the Lasso problem.**
  + The full name of the Lasso can be written as the Least Absolute Shrinkage and Selection Operator proposed by Tibshirani,
  + Here are some basic definitions defined in our setting, Note, lambda is the tuning parameter that control the strengthen of penalization, beta is the regression coefficient.
  + The lasso regression can be modelled as equation (1) and (2), where a l1 norm penalty is used for penalizing large coefficient. The lasso estimator can be obtained via equation 2. We will use the notation throughout this presentation.
* **Lasso: Importance and Shortage (Slide 5)**
  + The reason why we are introducing the Lasso problem is because of the following.
    - First, it can automatically shrink the insignificant coefficient, to achieve feature selection, as shown in figure 1, this picture demonstrates the contour plot between lasso regression and ridge regression that use l2 penalty instead of l1 penalty in lasso. The lasso regression tends to have a solution that lies in the axis, as opposed to the ridge regression that produces circled contour plot.
    - In addition, due to the sparsity nature of the model, it can prevent overfitting and it is equivalent to the l-1 regularization.
  + Despite that, there are also some limitations for the Lasso,
    - Importantly, it can’t capture the variance of the inferential quantity, this means the ordinary lasso can only capture an estimated regression coefficient.
    - Finally, there does not exist reliable and fast method for obtaining suitable or optimal lambda.
* **Bayesian Lasso: transition from the Ordinary Lasso (Slide 6)**
  + To address the issue of ordinary lasso, a Bayesian framework is embedded into the lasso framework. In particular, the use of a laplace prior to equation 3 is fitted into the model.
  + In addition, Park and Casella further extend the Bayesian Lasso by treating the regression coefficients as random variables and assigning them a conditional prior distribution as equation 4 to ensure unimodality for easy optimization and inference. This can help us to formulate the Bayesian Lasso more easily.
  + The Bayesian lasso also introduces an auxiliary variable \tau, and it can be integrated to return to the ordinary lasso setting as illustrated in equation 5.
* **Bayesian Lasso Formulation (Slide 7)**
  + Here we have given the hierarchical representation presented by the park and Casella. This hierarchical representation enables sparsity and variable selection while accounting for uncertainty in the estimation process. It uses a representation of the double exponential distribution as a mixture of normal, with the independence of tau square and sigma square.
* **Bayesian Lasso: Three-step Gibbs Sampler (Slide 8)**
  + The traditional Bayesian posterior is obtained by Gibbs Sampler mentioned in Park and Casella. Gibbs sampling iteratively samples conditional distributions of each parameter, with a long-time convergence, it will be arbitrarily close to the target distribution. Below is the derivation under the hierarchical Bayesian lasso setting.

Beta: is following a Multivariate Normal Distribution

Sigma square us following a Inverse gamma distribution.

Tau\_j^2 is following a Generalized Inverse Gaussian Distribution

* **Bayesian Lasso: Comparison (Slide 9)**
  + Here we’ve listed the advantages and disadvantages of the Bayesian Lasso
  + There are several advantages for the bayesian lasso, such as improved coefficient estimates, and better prediction accuracy. It can also achieve more reliable uncertainty quantification compared to the standard lasso, as well as the automatic selection of the regularization parameter: lambda, where the tuning parameter can be set as another variable by sampling.
  + NEVERTHELESS, the fundamental limitations of the Gibbs sampler for Bayesian Lasso are that it can be particularly costly when running. It can take a long time to converge.
* **Bayesian Lasso: MFVB (Slide 10)**
  + To address the long-time running issue for the Gibbs Sampler, a common alternative method is Mean Field Variational Bayes.
  + It assumes that the approximation originates from an analytically tractable class of distribution Q, with a form of factorization restriction as shown in equation 6
  + Afterward, it attempts to search for the distribution from this family that is closest to the target posterior distribution with some discrepancy metric as equation 7, such as the KL divergence given by equation 8. An optimization-based system is established by iteratively updating variational parameters with an appropriate optimization algorithm. A traditional optimization algorithm can be Coordinate Ascent algorithm, while the most common choice of Q is the Normal Distribution due to its simple form and adaptability to other distributions. And the optimal solution can be derived from Equation 9.
* **Bayesian Lasso: MFVB Update Procedure (Slide 11)**
  + This page is basically about how to update the posterior distribution parameter for the Bayesian Lasso under the mean field restriction framework.
  + If we derive the updated formula from the optimal solution from Equation 9
  + Then The update of beta leads to the update formula of the mean and sigma for Gaussian Distribution.
  + Also, the update of the sigma square leads to the updated formula for the shape and rate parameter of the Inverse Gamma distribution
  + Our final output will be the parameter for the posterior distribution as: tilde mu, Tilde sigma, \tilde a and tilde b
* **Bayesian Lasso: MFVB Compare with MCMC (Slide 12)**
  + Compared with MCMC, Variational Approximation is a hundred times faster, and it also has a lower computational cost.
  + Nevertheless, VI can suffer from low approximation accuracy and underestimate the posterior variance, if the predictors have a high correlation.
* **Proposed Algorithm Definition (Slide 13)**
  + **To further increase the approximation accuracy when MFVB accuracy is low, we would** like to explain more about our algorithm, as the name is local-global algorithm, it conducts a local approximation for each variable at first, followed by a global propagation to correct the final result.
  + Below are the necessary definitions for understanding our algorithm. Firstly, Local Approximation: focus on the approximation of the marginal posterior distribution.
  + Global Approximation: focus on approximating the joint posterior distribution. For example: they will formulate a multivariate distribution.
  + Below is the comparison between our method and the MFVB method?
    - MFVB will produce global approximation only.
    - But Our algorithm will use the information from the local approximation to correct the global approximation.
* **Proposed Algorithm Basic Setting (Slide 14)\**
  + **Now Let’s move to the core intuition of our algorithm.**
  + Similar to Variational Inference, the mean Field assumption will be used again. So, equation 10 states that target distribution is approximated to q(\beta) times q(\sigma^2),
  + The initial input will be based on MFVB output, the lambda is also generated by posterior mean from Gibbs Sampler for now.
  + In addition, since we don’t have enough time to explore the derivation of sigma square, thus we are assuming independence of beta and sigma^2,
  + Finally, our goal is to find the optimal mu and sigma square so that get the best approximation to the target posterior distribution.
* **Local Likelihood Derivation (Slide 15)**
  + Next, we will introduce briefly about our local approximation adjustment,
  + We introduce a conditional distribution setting for variational inference as equation 11, where an expectation with respect to the conditional distribution q(beta minus j given by \betaj) is used to get the Evidence Lower bound.

In contrast to the MFVB, MFVB assumes independence between beta minus j and beta j, while our method could capture the correlation between beta minus j and beta j.

* + In the Bayesian lasso problem, the marginal log-likelihood can be derived and written in equation 12, which is proportional to a univariate lasso distribution with a corresponding parameter as stated in equation 12.
* **Lasso Distribution: pdf (Slide 16)**
  + Before we further explore our algorithm, we will first introduce our newly invented distribution: Lasso distribution,it is in the form of an exponential function as Equation 13.
  + The pdf can be divided up into four components.
    - A normalization constant, to make the integration of this pdf to 1.
    - A quadratic term ax square to control the curvature of the curve
    - A linear term bx to control the location of the curve,
    - An absolute term c times absolute of X to control the sharpness of the turning point.
* **Lasso Distribution: Graphical Illustration (Slide 17)**
  + **To provide a better understanding of the shape of the Lasso Distribution, here is the visualization of the lasso distribution given 6 parameter settings.**
  + From the yellow and dark-blue line, we can observe that parameter A control the size of the curvature of the tuning point, larger A implies a smoother tuning point.
  + From the Red line, yellow line, and sky-blue line, we can observe that changing b will move the location of the curve. Larger b will move the graph further to the right
  + From the yellow line and green line, C controls the sharpness the curve, larger c implies distribution with smaller variance and shaper turning point
* **Lasso Distribution: Properties 16 (Slide 18)**
  + We have also derived the corresponding property for a probability density function, for example, the normalizing constant can be written in the form of the first bullet point, and the moment can be written in a mixture sum of products based on the Positively Truncated normal distribution, and variance can be derived by the variance formula.
* **Local Approximation Adjustment (Slide 19)**
  + CONTINUING introducing our algorithm, we know the marginal likelihood can be approximated by a lasso distribution, this property helps us to better calculate the local mean mu j star and local variance Sigma jj star as a new input to a new marginal Gaussian Approximation so that provide a more exact local approximation.
  + Moreover, the conditional distribution q (beta minus j given by beta j) for any jth variable can be derived by the fact that q(\beta) is assumed to be from a normal distribution. Therefore, q (beta minus j given by beta j) can be written in a multivariate normal distribution from Equation 16
* **Global Approximation Propagation (Slide 20)**

After calculating the local mean, local variance, and conditional distribution q, the global parameter can be propagated through equation 17. The final parameter update formula can be written in equation 18 and equation 19 respectively. It is expected that the mu tilde and sigma tilde will be the best approximation for the posterior of beta.

* **Univariate Local-Global Algorithm (Slide 21)**
  + The local and global algorithm can be mainly divided up into three steps given by the following pseudo-algorithm.
    - Firstly, it will calculate the current a, b, and c based on current data.
    - Secondly, it will calculate the local mean and local variance of the marginal distribution, by a lasso mean and lasso variance
    - Thirdly, it will propagate using equation 18 and equation 19 to correct global parameters.
* **Experiment: setup: dataset description 20**
  + To examine the effectiveness of our algorithm, we’ve conducted experiments on 6 datasets,
  + Here is the statistic of the dataset, the most evident dataset that needs to be paid attention to are Hitters and Eyedata, where Hitters is a Baseball statistics dataset that shows a high correlation between predictors, and Eyedata is a Medical dataset that have higher predictors than number of samples, which means they are harder to approximate.
* **Experiment: setup: evaluation metric 21**
  + We’ve also used some evaluation metrics to examine the effectiveness of our algorithm
  + L1 norm accuracy is defined by equation 20 and equation 21,
    - this accuracy is concentrating on the accuracy in the center of the distribution, rather than the tailed distribution.
  + In addition, the execution time of the algorithm can help us measure the time complexity of the algorithm and compare it with other algorithms such as MFVB.
* **Experiment: Experiment result: approximation accuracy 22-23**
  + **Table 2 shows the experiment results for the average approximation accuracy of 3 different algorithms on 6 datasets. LG\_local represents the local approximation accuracy for our algorithm and LG\_global represents the global approximation accuracy in our algorithm.**
  + **From Table 2, we can observe that LG\_local surpasses all other accuracy except for the MCMC, which is the gold standard**
  + **Table 3 shows the average approximation speed experiment results for the 3 different algorithms of 6 datasets. Again, even though the Local global algorithm shows a slower speed than MFVB, it is a significant improvement in running speed than MCMC**
* **Experiment: Experiment result: approximation density visualization 26-27**
  + The figure 3 shows an approximation density plot for different algorithms, this plot shows one of the worst approximation density plots on the Hitters dataset. The green one shows MCMC, the blue is MFVB, the red is a local and global algorithm to local approximation and the purple is a local and global algorithm to the global approximation. As we can see from the result in this particular predicor, MFVB tends to deviate from MCMC (gold standard), especially when the actual distribution has a sharp tuning point, the local and global algorithmshowever can approximate well to the gold standard.
  + **25**
  + Figure 4 shows the result in Eyedata, we can get a similar conclusion as before.
  + In addition, theoretically, the approximation of local approximation is better than global, it is also expected as most of the MCMC plot overlap with the LG\_local plot.
  + It is more evident to observe there are more discrepancies in Eyedata than in Hitters.
* **Experiment: Discussion 26**
  + From the experiment result, we can obtain some conclusions:
    - Firstly, MFVB tends to generate a density plot with a higher variance sometimes, causing a discrepancy in accuracy. This is mainly because of the product factorization restriction globally.
    - On the other hand, local-global algorithm demonstrates a closer density plot posterior distribution in all cases.
    - In general, our algorithm produces a better result in approximation accuracy with descent speed.
* **Limitation and Future work 27**
  + Due to limited amount of time, there are some improvements awaiting to be explored in the future.
  + The first limitation is that the automatic choice of λ is now currently obtained by Gibbs Sampler with long time execution time, we can develop automatic selection methods based on our algoirhtm.
  + Secondly, more evaluation metrics such as matrix norm can be used to understand the performance of our algorithm comprehensively.
  + Finally, if the initial covariance is diagonal, by our update formula, it will remain diagonal for the last iteration, which makes the univariate local-global algorithm non-generalizable.
  + In the future, we can introduce another algorithm called Bivariate local and global algorithm, that match each pair of variables to form a bivariate lasso distribution locally and propagate to the global parameter.
  + Also, the update formula for Sigma can also be further derived and examined later.
* **Conclusion 28**
  + In conclusion, this project has achieved success in exploring a new way to enhance Bayesian Lasso Posterior distribution approximation, providing a higher approximation accuracy.
  + In addition, the time of execution is slower than MFVB since MFVB has only global updates.
  + We’ve also invented a lasso distribution for better fitting the lasso posterior marginally. Its corresponding property such as expectation and covariance has also been derived and implemented in R.
  + In the future, we will address the issue mentioned in limitation on the last slide and try to generalize our algorithm to more problem in the future.
* **Acknowledgement**
  + Finally, I would like to express my deep thanks to my supervisor John Ormerod and Co-supervisor Doctor Mohammad Javad Davoudabadi, for their constant guidance and patience throughout this year.
  + Here is the reference used in this presentation, Thank you for listening!
* QA
  + Why method?
    - Slide 13
    - Slide 15: if we look at equation 11, our method take expectation with respect to a conditional distribution q(\beta minus j given by beta j),
    - While MFVB will take expectation with respect to distribution q(\beta)
    - MFVB also assume independence between marginal distribution, while ours capture more correlation between targets parameter
    - This is also from the fact that if t=0, then s = mu minus j
  + Why MFVB output as input
    - To be more accurate, due to deficiency of univariate local-global algorithm