NORMALIZING FLOWS AIDED VARIATIONAL INFERENCE FOR UNCERTAINTY QUANTIFICATION

By

Sumegha Premchandar

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Statistics – Doctor of Philosophy

2024

ABSTRACT

Bayesian statistics is a powerful tool for quantifying uncertainties when estimating unknown model parameters. It is often the case that the posterior distributions arising from the Bayesian paradigm are intractable. This may be due to complex statistical model choices and high-dimensionality of the parameter space. Previously, Markov Chain Monte Carlo (MCMC) methods have been the preferred approach for sampling from posterior distributions with an unknown normalizing constant. However, MCMC methods run into a number of issues in practice. For instance, they do not always scale well to multimodal distributions defined on a high-dimensional support. Variational Inference (VI) has emerged as a scalable alternative to MCMC for sampling from intractable posterior distributions. Recently, Normalizing Flows aided VI (FAVI) has been used for sampling from complex and multimodal posterior distributions to overcome the limitations of existing mean-field and structured VI approaches. FAVI has had a significant impact across fields in applications such as computer vision, computational biology, and physics-based modelling. Despite its impact, there is limited research on the theoretical properties of the approximate posterior arising from FAVI. The computational cost of FAVI depends heavily on the choice of Normalizing Flow (NF) family, but there is no work quantifying the nature of the approximate posterior from FAVI at a particular complexity of the NF, especially with respect to uncertainty quantification.

In this dissertation, we study the properties of the FAVI posterior with a focus on:

(i) The trade-off between accurate recovery of the posterior samples and complexity of the selected NF family.

(ii) Uncertainty quantification.

We first provide background on FAVI and compare it to popular competitors (Mean-Field VI (MF-VI) and MCMC) over some basic statistical applications. Our results demonstrate that FAVI lies between MCMC and MF-VI in both statistical accuracy and computational efficiency.

In this second part of this dissertation, we use the framework of Bayesian linear regression with 2 predictor variables to rigorously study the optimal Kullback-Leibler divergence between

the FAVI approximation with Inverse Auto-regressive Flows (IAF) and the true posterior. We also derive the uncertainty quantification (credible interval coverage) resulting from using FAVI to approximate the posterior, as a function of the correlation between the regression predictors. We contrast this coverage with MF-VI (the most popular VI approach in the literature) and find that, given sufficient complexity of the NF, there is virtually no loss in coverage from FAVI relative to the true posterior, regardless of the correlation. On the other hand, the loss in coverage for MF-VI increases monotonically in the correlation.

Next, we extend our results to the case of an arbitrary $p > 2$ regression predictors. Our results (presented across complexity levels of the IAF transformations), demonstrate that given sufficient complexity of IAF, FAVI can completely recover the true posterior. To our knowledge, this is the first theoretical exploration of this kind.

Finally, we discuss ongoing research and plans for future work where we will leverage our learning to use FAVI for Bayesian inference in high-dimensional linear models with spike and slab priors. Preliminary results show that FAVI can capture dependencies in the posterior more effectively than MF-VI.

FAVI is one among many novel computational tools that has originated in machine learning literature for scalable Bayesian computation, but there has been little previous work analyzing its statistical properties and reliability for uncertainty quantification. By studying the FAVI posterior from a statistical lens, this dissertation bridges some of the gap between machine learning and statistics, and takes strides towards building reliable computational tools for Bayesian inference.

This thesis is dedicated to my parents - Jayanthi Premchandar and M.R. Premchandar.

ACKNOWLEDGEMENTS

First and foremost, I would like to express my gratitude to my advisors Dr. Tapabrata Maiti and Dr. Shrijita Bhattacharya who have been instrumental in shaping my thesis. Their enthusiasm for tackling interesting and challenging research problems has been an example for me and has molded me into the researcher I am today. I have greatly appreciated their support and patience these past 3 years while I learned the ins and outs of doing research. I would also like to express my gratitude for the generous funding provided to me by my advisors during my third and fourth years of the PhD. This funding gave me the space and time to focus on my research. I am additionally very thankful to my committee members Dr. Yimin Xiao and Dr. Selin Aviyente for the time they spent attending committee meetings, reading my work and the valuable feedback they provided that greatly improved the quality of my work.

I count myself lucky to have had some wonderful professors and mentors over the course of my PhD including Dr. Yimin Xiao, Dr. Haolei Weng and Dr. Shrijita Bhattacharya whose instruction in the prelim courses helped me build a strong foundational knowledge in probability and statistics. I am additionally grateful for the mentorship of Dr. Sandeep Madireddy for providing me with the opportunity to intern at Argonne National Laboratory and for introducing me to the exciting area of Neural Architecture Search. My time at Argonne was well spent, picking up computational and research skills that have been put to good use throughout my PhD. Aside from my professors and mentors, I would like to express my appreciation for the STT staff; particularly Andy, Tami and Ashlynn for making administrative processes far easier and MSU technology much more accessible.

I have had a small but loyal army of friends and family by my side who have made this journey possible. To my friends Tathagata, Sikta and Hema; thank you for providing me with a sense of community and family in East Lansing. Thanks also to my other friends at MSU; Phuong, Satabdi, Soyeong, Nian, Haoxiang, Sang Kyu, Arka, Anirban, Sampriti, Alex and Sanket for making East Lansing feel much less lonely. I am especially indebted to Arka, who allowed me to pick his brain about my research many times and who has always been a source of useful advice for me.

My support system outside of STT has been just as important as the people inside of it. I greatly

value my family members, who have provided an unwavering love and support for me these past few years. They have made the effort to keep in touch even when my busy schedule sometimes left me with little energy to do the same. I reserve a special mention for all of my grandparents, who have been proud of even the smallest of my wins. Thanks to my friends Mahevash and Shailja for coming all the way to Chennai to visit me on my first trip back to India and for our enduring long-distance friendship. I owe a great deal to my friends Anisha and Nayantara; who have always been there for me at my lowest points and for reminding me of my worth outside of research. It is difficult for me to find the words to describe how much their friendship has meant to me. The near daily pictures of Ingee, my favourite ginger cat, and her sage advice communicated via Anisha kept my spirits amused on many an occasion.

I can state with certainty that I would not have made it to the finish line without the love and support of my partner Nisarg. He has helped me with the more frustrating parts of using LaTeX, debugged my code on many occasions and acted as a sounding board for my ideas. But more than this, thank you for having faith in me when I had none in myself, for encouraging me to treat myself more kindly and bringing so much levity to even the most stressful parts of my life. My deepest appreciation is for my sister Sucharita, whose unconventional and bite-sized wisdom got me through the more arduous phases of this PhD. Last and most important, I would like to thank my parents Jayanthi Premchandar and M.R. Premchandar for everything they have done for me. Thanks Appa, for doing the sometimes difficult job of pushing me when it was required and thank you Amma for being such a source of strength, resilience and unconditional love for me these past 5 years and more. This thesis belong to all of you.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

BVS	Bayesian variable selection
CDF	Cumulative distribution function
ECDF	Empirical cumulative distribution function
FAVI	Flows aided variational inference
GLM	Generalized linear model
IAF	Inverse auto-regressive flow
KDE	Kernel density estimates
MCMC	Markov chain Monte Carlo
MF	Mean-field
MH	Metropolis Hastings
NAF	Neural auto-regressive flows
NF	Normalizing flows
PDF	Probability distribution function
PMF	Probability mass function
RW	Random-walk
SGA	Stochastic gradient ascent
SSE	Sum of squared error
SVI	Structured variational inference
VI	Variational inference
DSF	Deep sigmoidal flow

CHAPTER 1

INTRODUCTION

In various scientific fields, statistical and machine learning models play a significant role in inference and decision making. It is crucial to have models that accurately represent uncertainties in unknown model parameters, as this helps in building robust decision making processes. This is especially important in safety critical applications such as medical image analysis and autonomous driving. Bayesian statistics is a potent tool for quantifying uncertainties, without needing to resort to complex asymptotic results. It also has the added advantage of being able to incorporate domain specific prior knowledge into the statistical model.

Bayesian statistics derives all inference about an unknown model parameter from the posterior distribution. In numerous applications, the posterior distribution does not have a closed form. In such a situation we say that the posterior distribution is "intractable" and we use approximate inference methods to sample from it. A key challenge in Bayesian statistics is developing scalable and statistically accurate computational tools to generate samples from complex, intractable posterior distributions. It is difficult to balance the trade-offs between statistical accuracy and computational efficiency for approximate inference methods.

Normalizing Flows aided Variational Inference (FAVI) is an algorithm for sampling from intractable distributions that originated in machine learning literature [33]. It was introduced to recover complex and multimodal distributons, while retaining some of the scalability that characterizes VI. FAVI been impactful across scientific fields such as computer vision [20], computational biology [7] and physics-based modelling ([43], [46]). Given its popularity across application areas, it is crucial to address some of the fundamental theoretical gaps in our knowledge surrounding this valuable tool. This can then enable its wider adoption for statistical inference and variable selection in high-dimension.

In the rest of this chapter we will provide an overview of concepts relevant to our work and discuss important related literature. At the end of this chapter, we provide an outline for the rest of this dissertation.

1.1 Markov Chain Monte Carlo

Markov Chain Monte Carlo (MCMC) methods generate samples from a Markov Chain with a stationary distribution equal to the target posterior distribution we wish to sample from. There are a number of MCMC methods in the literature, the most well known of which are the Metropolis Hastings algorithm [9] and Gibbs sampling [8].

The Metropolis-Hastings algorithm uses a proposal distribution that serves as a transition kernel for a Markov Chain. The proposal distribution is used to generate samples (a new state) based on some previous state. This new state is then accepted or rejected with a probability that depends on the un-normalized target posterior and the proposal distribution. One of the most popular choices for the proposal is a Gaussian distribution with mean equal to the previous state. This is known as the Gaussian Random-Walk Metropolis Hastings (RW-MH) algorithm and we use it for our comparison studies in chapter 2.

Gibbs sampling is a special case of the Metropolis-Hastings algorithm. Gibbs sampling generates samples from a multivariate distribution, one dimension at a time. If we have a joint distribution $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_p) \in \mathbb{R}^p$, it leverages information about the complete conditional distributions $\pi(\theta_i|\theta_1, \theta_2 \dots \theta_{i-1}, \theta_{i+1}, \dots \theta_p)$, $1 \leq i \leq p$, to produce samples.

MCMC methods have the desirable property that the generated Markov chain samples are guaranteed to converge to samples from the target distribution. However, they run into a number of issues in practice. MCMC can be slow to converge, especially for high-dimensional state spaces and multimodal target distributions. The RW-MH algorithm requires a computational budget of $O(p^2)$ to generate two approximately independent samples from the stationary distribution [26]. If a large number of samples are required, this can be computationally prohibitive. Gibbs sampling generally has faster mixing times than the RW-MH algorithm and does not require tuning of the proposal distribution, but we may not always have information on the complete conditionals.

There are more contemporary MCMC methods such as the Hamiltonian Monte Carlo (HMC) that leverage gradient information of the target distribution to explore the state space much more efficiently than the MH algorithm. However, as discussed later in section 2.7, even HMC runs into

issues for highly multimodal target distributions.

Lastly, assessing convergence of the Markov Chain to the stationary distribution is often challenging [35]. Empirical convergence diagnostics such as the Gelman-Rubin diagnostic [10] are popularly used, however they do not always guarantee convergence of the Markov Chain. Further, calculating such diagnostics requires running multiple parallel chains from which many of the samples are discarded. This can be computationally expensive. Other measures such as upper bounds on the total variation distance between the density of the MCMC samples and the stationary distribution are difficult to obtain for many statistical models.

1.2 The basic Variational Inference algorithms

Variational Inference (VI) surfaced in machine learning literature as a scalable means of sampling from intractable posterior distributions ([18]). VI is widely used in applications such as computer vision, topic modeling, and computational biology ([5], [7]). A common theme in these applications is the presence of large datasets and high-dimensional parameter spaces. In VI, a variational family of distributions Q is proposed, from which a distribution q_{ϕ^*} is selected based on its "closeness" to the target posterior. Keeping with the formulation in a majority of VI literature, we will use the Kullback-Leibler (KL) divergence as the measure of closeness. More precisely, VI reframes the problem of sampling from a target distribution $\Pi(\boldsymbol{\theta}|D)$ into the following optimization:

$$q_{\phi^*} \in \arg\min_{q_\phi \in Q} KL(q_\phi \,||\, \Pi(.|D)) \tag{1.1}$$

The choice of family Q drives the trade-off between statistical accuracy and computational efficiency that characterizes VI. Mean-Field VI (MF-VI), enables computational efficiency by assuming any $q_\phi \in Q$ can be factorized as $q_\phi(\boldsymbol{\theta}) = \prod_{i=1}^{p} q_{\phi_i}(\theta_i)$. Structured VI (SVI) allows for some level of dependencies among θ_i by estimating a non-diagonal covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$, but it sacrifices some of the computational efficiency of MF-VI. A caveat of both MF-VI and SVI is that they cannot recover multimodal target distributions.

1.3 Normalizing Flows aided Variational Inference

Some of the earliest mentions of the idea of using a Normalizing Flow for probabilistic modelling can be found in [39] and [38]. A Normalizing Flow (NF) is nothing but a composition of continuously differentiable mappings with differentiable inverse, applied to samples from a base distribution. **N**ormalizing **F**lows aided **VI** (FAVI) was introduced in [33] to alleviate some of the issues encountered by MF-VI and SVI in sampling from complex and multimodal target distributions. FAVI generates a family of distributions by sampling from a base distribution $q_0(\boldsymbol{\theta}_0)$ and applying differentiable and invertible mappings $T_s : \mathbb{R}^p \to \mathbb{R}^p$ such that $\boldsymbol{\theta}_S = T_S \circ T_{S-1} \cdots \circ T_1(\boldsymbol{\theta}_0)$. If $\varphi \subseteq \mathbb{R}^m$ denotes the space of parameters for the transformations $\{T_s\}_{s=1}^S$, then any $\phi \in \varphi$ results in a distribution $q_\phi(\boldsymbol{\theta}_S)$. We then choose an optimal distribution q_{ϕ^*} based on the minimization in (1.1). More complex choices of T_s yield more expressive variational families, albeit at added computational cost. Among the many ways to choose T_s, we limit our scope throughout this dissertation to the popular Auto-Regressive Flows, for which the cost of computing $q(\boldsymbol{\theta}_S)$ from $\boldsymbol{\theta}_0$ is $O(pS)$ [28].

FAVI has demonstrated a lot of potential as a computational tool for Bayesian inference in complex statistical models ([28], [22]). However, very little is known about theoretical behaviour of the variational posterior arising from FAVI. Previous theoretical works on VI have mostly focussed on asymptotic posterior consistency properties of the approximate posterior arising from MF-VI or SVI ([3], [31], [42]). These results take a frequentist viewpoint and are focussed more on the impact of the variational approximation on central tendency estimates.[1] They do not provide an in-depth exploration on the uncertainty quantification obtained from the variational posterior. Certain families of NFs are known to be highly expressive and in theory, they can model any target distribution with non-zero support on $\mathbb{R}^p$ (see chapter 2). Given these properties, FAVI should demonstrate improved uncertainty quantification and recovery of the posterior samples for finite sample sizes, when compared to simpler variational families (e.g. mean-field). Consequently, we believe that while studying the theoretical properties of the approximate posterior obtained from

[1] By frequentist viewpoint we mean the assumption that there exists a true unknown parameter $\boldsymbol{\theta}_0$, that generates the observed data D, by means of a probability distribution $\mathbb{P}(D|\boldsymbol{\theta}_0)$.

FAVI, it is crucial to look beyond frequentist consistency results to assess the overall statistical accuracy of posterior samples from a Bayesian perspective, especially with respect to uncertainty quantification.

Chapters 2 and 3 provide more background on FAVI, as well as a current state of research in the area, open problems and necessary technical details.

1.4 Preliminary Notation

We use lowercase letters to denote scalars ($x \in \mathbb{R}$), boldface letters to denote vectors ($\boldsymbol{x} \in \mathbb{R}^m$) and capital letters denote matrices ($X \in \mathbb{R}^{m \times d}$). The symbols $\Phi(.)$ and $\phi(.)$ represent the standard normal cumulative distribution function (cdf) and probability distribution function (pdf) respectively. The notation $\perp\!\!\!\perp$ is used to represent pair-wise independence between any two random variables. Let $\mathbb{I}$ denote an indicator function, that is, for any real valued random variable Y and Borel set $A \in \mathcal{B}(\mathbb{R})$; $\mathbb{I}(Y \in A) = \begin{cases} 1, & \text{when } Y \in A \\ 0, & \text{otherwise} \end{cases}$

For any scalar $s \in \mathbb{R}$ we use $\text{sign}(s)$ for the sign of s, that is:

$$\text{sign}(s) = \begin{cases} 1, & \text{if } s > 0 \\ -1, & \text{if } s < 0 \\ 0, & s = 0 \end{cases}$$

The size m identity matrix is given by $I_m \in \mathbb{R}^{m \times m} = \begin{bmatrix} 1 & 0 & \dots 0 \\ 0 & 1 & \dots 0 \\ \vdots & \vdots & \vdots \\ 0 & 0 & \dots 1 \end{bmatrix}$.

We use the symbol Q to denote a variational family of distributions and q_ϕ is a member of Q with variational parameters ϕ. We use q_{ϕ^*} to refer to the variational posterior, that is the optimal distribution from Q that minimizes the KL divergence as per equation (1.1). We will sometimes use q_{ϕ^*} to refer to a specific member of Q that may not satisfy the minimization (1.1), but we will make it clear if that is the case.

The symbol $\boldsymbol{\theta}$ is used to denote the unknown parameter of interest or latent variables and D refers to observed data. We will use p to denote the dimensionality of the parameter space ($\boldsymbol{\theta} \in \mathbb{R}^p$) and $\Pi(\boldsymbol{\theta}|D)$ or $\Pi(.|D)$ for the target posterior distribution. The Kullback-Leibler (KL) divergence between 2 probability distributions q and p is defined as:

$$KL(q \mid\mid p) = \int_{\boldsymbol{x}} \ln \frac{q(\boldsymbol{x})}{p(\boldsymbol{x})} q(\boldsymbol{x}) d\boldsymbol{x}$$

See Tables 3.1 and 3.2 for more details on mathematical notation used in this dissertation.

1.5 Dissertation Outline

In this dissertation we study the properties of the variational posterior arising from FAVI with a dual focus on: (i) the trade-off between accurate recovery of the posterior samples and complexity of the NF transformations T_s (ii) uncertainty quantification.

Chapter 2 serves as an exposition of FAVI and is written to be accessible to a broad scientific audience. It also includes comparisons to popular competitors (Mean-Field VI (MF-VI) and MCMC) over a variety of classical statistical applications. The results of our comparison studes indicate that FAVI lies somewhere between MCMC and MF-VI in statistical accuracy and computational efficiency. This motivates the problems we consider in the subsequent chapters of this dissertation.

In Chapter 3 we begin with a rigorous study of the optimal Kullback-Leibler divergence and loss in uncertainty quantification between the FAVI approximation with Inverse Auto-regressive Flows (IAF) and the true posterior, within the specific context of Bayesian linear regression with 2 predictors. We also contrast the loss in uncertainty quantification (credible interval coverage) from using the FAVI approximate posterior with IAF to that of MF-VI as a function of the correlation between regression predictors. We then extend our results to the case of $p > 2$ regression predictors. Our theoretical results highlight the benefits of FAVI for uncertainty quantification.

We follow this with details on ongoing research, where we adapt FAVI for Bayesian inference in high-dimensional linear models, with spike and slab priors (chapter 4). We demonstrate its usefulness in capturing dependencies across latent variables as well as emphasize its limitations. Based on these limitations we suggest possible directions for future work.

CHAPTER 2

NORMALIZING FLOWS AIDED VARIATIONAL INFERENCE: BACKGROUND, EXAMPLES AND COMPARISONS

A modified version of this chapter was first published in Notices of the American Mathematical Society in volume 70, number 7, year 2023; published by American Mathematical Society.

2.1 Background

A major area of contemporary statistics research is learning to model probability distributions of varying complexity. The problem of learning to characterize probability distributions broadly takes 2 forms: estimating a probability density given samples from it and approximating densities that are known only up to a normalizing constant. The latter avenue of research has applications in Bayesian inference, where we wish to generate samples from the posterior distribution of model parameters given observed data.

This chapter discusses the use of Normalizing Flows for Variational Inference (VI), a method wherein we can approximate and sample from complex probability densities [33]. This type of probabilistic modeling lies in the second avenue of research, where we do not have a normalizing constant for probability densities of interest. VI is a tool that emerged in machine learning to approximate probability densities. It is often applied in Bayesian statistics as a more scalable alternative to Markov Chain Monte Carlo (MCMC) methods for large datasets. Although scalable, earlier works such as mean-field or structured VI are limited when approximating more complex and multimodal probability distributions. Normalizing Flows are mappings from a simple base distribution to a more complex probability distribution. They are primarily used for modeling continuous distributions and can be used to specify very flexible probability models, thus improving the statistical accuracy of VI algorithms.

There already exist comprehensive reviews for Normalizing Flow methods in general. An overview of different Normalizing Flow families is provided in [22], while [28] goes into depth on each family of flow models and extends this discussion to newer areas, such as flows for discrete

variables. These reviews are an overarching look at flows for probabilistic modelling and are focussed on applications in the machine learning literature. Discussion of applications of a more classical statistical nature is limited. An excellent exposition and survey of VI from a statistical lens is given in [4]. However, they only cover variational families with a known parametric form, such as mean-field and structured VI. We extend the discussion to variational families specified by Normalizing Flows. Further, this chapter is written to be accessible to readers new to the area.

In latent variable modeling, we aim to learn the conditional distribution of latent variables $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_p)$ given observed data D, that is, $\Pi(\boldsymbol{\theta}|D)$.[1] We explain how solving this problem is useful in Bayesian statistics. In parametric statistics, stochasticity in the observed data is often described using a specific probability distribution $p(D|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ needs to be estimated based on the data D. In Bayesian inference, we assume a prior distribution $\pi(\boldsymbol{\theta})$ on $\boldsymbol{\theta}$ representing our beliefs about the model parameter *prior* to observing the data. Based on the data, we update our beliefs via the posterior distribution $\Pi(\boldsymbol{\theta}|D)$. The posterior can be calculated by Bayes theorem:

$$\Pi(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

For cases where the marginal likelihood $m(D) = \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is intractable we resort to approximate inference. MCMC methods have long been the go-to for sampling from posterior distributions when $m(D)$ cannot be computed. MCMC algorithms generate samples from a Markov Chain whose stationary distribution converges to the target distribution of interest. One prominent example is the Metropolis-Hastings method [9], of which the Gibbs sampling algorithm [8] is a special case. However, these methods may not always scale well to high-dimensional models and can be slow to converge for multimodal distributions. VI has shown promise as a scalable alternative to MCMC. In VI, the target distribution is approximated by a family of distributions Q among which we choose the optimal distribution q_{ϕ^*} to be "closest" to the target. To determine "closeness", KL-divergence is often used. Intuitively, KL-divergence is something akin to a distance between 2 probability distributions. Thus, probabilistic modeling with VI becomes an optimization

[1] In much of the variational inference literature, $\boldsymbol{z}$ will be used for the unknown parameters (latent variables) instead of $\boldsymbol{\theta}$. We use $\boldsymbol{\theta}$ to be consistent with statistics literature.

problem:

$$q_{\phi^*} \in \arg\min_{q_\phi \in Q} KL(q_\phi \,||\, \Pi(.|D))$$

Mean-Field VI (MF-VI) is a popular approach in which the variational family Q is defined based on the assumption that latent variables are independent of each other. The mean-field assumption is useful for faster computations during optimization but is restricted in the complexity of densities we can approximate. Structured VI takes this one step further by allowing dependencies across latent variables. However, even with Structured VI we cannot guarantee that we can approximate *any* density arbitrarily well. This is where Normalizing Flows come in.

2.2 When should we use Normalizing Flows VI?

In [4], the authors observe that "VI is suited to large data sets and scenarios where we want to quickly explore many models; MCMC is suited to smaller data sets and scenarios where we happily pay a higher computational cost for more precise samples." While this is generally true of MF-VI, Normalizing Flows VI lies somewhere between MCMC and other variational approximation approaches in terms of computational efficiency and accuracy. To shed some light on how Normalizing Flows VI compares to other sampling methods such as MCMC and MF-VI, we implement variational inference with Neural Auto-regressive Flows [16] for several examples. These examples cover classical Bayesian statistical applications in exponential family models, Gaussian linear regression and logistic regression. We cover scenarios of varying dimensions and complexity of the target distribution. This gives us a high-level idea of scalability vs. accuracy for these methods.

We begin the following section by introducing Normalizing Flows and elaborate on how to use them for VI. We then proceed to examples in Section 2.6. Finally, we discuss some important takeaways & challenges remaining in the area in Section 2.7.

2.3 Normalizing Flows

The main idea behind Normalizing Flows is to transform some simple continuous base distribution into a "target" distribution that is usually more complex, via a series of bijective, continuously differentiable transformations with differentiable inverse [28]. These functions are often referred

to as "diffeomorphisms".

Let $Z \in \mathbb{R}^p$ be a random variable whose density we wish to model. We begin with a random variable U sampled from some base distribution $p_U(\boldsymbol{u})$ defined on support $\mathbb{R}^p$ and apply a diffeomorphism $T : \mathbb{R}^p \to \mathbb{R}^p$ such that $Z = T(U)$. The density of Z is then given by the change of variable formula [36]:

$$p_Z(\boldsymbol{z}) = p_U(\boldsymbol{u})|J_T(\boldsymbol{u})|^{-1}$$

$|J_T(\boldsymbol{u})|$ denotes the absolute value of the determinant of Jacobian of T w.r.t $\boldsymbol{u}$. Thus, the function T transforms the density $p_U(\boldsymbol{u})$ into $p_Z(\boldsymbol{z})$. This process, wherein samples from one probability density 'flow' through a mapping to obtain another density is called a Normalizing Flow.

A natural question to ask is whether Normalizing Flows can be used to transform a simple base distribution (e.g uniform or standard normal distribution) into ***any*** target distribution. The paper [28] contains a constructive argument to show that Normalizing Flows can indeed recover any target density under rather general conditions. In practice, this is heavily dependent on the transformations T that we employ.

2.4 Discrete and Continuous-Time Flows

Normalizing Flows are mainly of two types - discrete time (finite flows) and continuous time (infinitesimal flows) [28]. Discrete-time Normalizing Flows are constructed by choosing a finite sequence of transformations $T_1, T_2, \ldots T_S$ and applying them successively to some base distribution $p_U(\boldsymbol{u})$ such that $\boldsymbol{z}_S = T_S \circ T_{S-1} \cdots \circ T_1(\boldsymbol{u})$. Since we choose all transformations to be diffeomorphisms, the change of variables formula applies and we have:

$$\begin{aligned} p_{Z_S}(\boldsymbol{z}_S) = p_U(\boldsymbol{u}) &\times |J_{T_S}(\boldsymbol{z}_{S-1})|^{-1} \\ &\times |J_{T_{S-1}}(\boldsymbol{z}_{S-2})|^{-1} \cdots \times |J_{T_1}(\boldsymbol{u})|^{-1} \end{aligned}$$

The number of transformations S, is often called the flow depth. Increasing flow depth can help us model progressively more complex densities at the expense of increased computational cost due to the calculation of $J_{T_s}(\boldsymbol{z}_{s-1})$.

We can think of discrete time flows as modelling the evolution of a probability density at S-many time points [28]. In contrast, continuous-time Normalizing Flows model this evolution continuously from some time $t = 0$ to S as an ordinary differential equation $\frac{d\boldsymbol{z}_t}{dt} = f(t, \boldsymbol{z}_t)$. A well known example of a continuous time flow is the Hamiltonian Flow, which is used for MCMC sampling [30].

2.4.1 Normalizing Flows for Variational Inference

We now expand on how Normalizing Flows are used to aid VI. We revert to our previous notation of using $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots \theta_p)$ to represent the latent variables, D for the observed data and $\Pi(\boldsymbol{\theta}|D)$ for the target conditional distribution we wish to sample from.

$$\Pi(\boldsymbol{\theta}|D) = \frac{p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{m(D)}$$

Recall that, VI approximates the target distribution by choosing a family of distributions $Q = \{q_\phi | \phi \in \Phi\}$ and selecting the optimal distribution in this family q_{ϕ^*} closest to the target density in terms of KL-divergence:

$$q_{\phi^*} \in \arg\min_{q_\phi \in Q} KL(q_\phi \,||\, \Pi(.|D)) \tag{2.1}$$

Other metrics such as more generalized α-divergence measures [24] can be used in place of KL-divergence. However, KL-divergence is popular due to its versatility and relative ease of implementation. The optimization in (2.1) is difficult to work with due to the presence of the intractable marginal likelihood $m(D)$. In practice, we maximize the **E**vidence **L**ower **Bo**und (**ELBO**) with respect to the variational parameters ϕ due to its equivalence to (2.1). The **ELBO** is the negative KL-divergence between the variational distribution q and the joint distribution $p(D, \boldsymbol{\theta})$ of latent variables and the observed data.

$$\begin{aligned} &\max_{q_\phi \in Q} \mathbf{ELBO}(q_\phi \,||\, \Pi(.|D)) \\ &= \max_{q_\phi \in Q} \Big\{ \mathbb{E}_{q_\phi(\boldsymbol{\theta})} \big[\ln p(D, \boldsymbol{\theta}) \big] - \mathbb{E}_{q_\phi(\boldsymbol{\theta})} \big[\ln q_\phi(\boldsymbol{\theta}) \big] \Big\} \end{aligned} \tag{2.2}$$

Using Normalizing Flows to aid Variational Inference was first popularized in [33]. The idea is to start with some base distribution $q_0(\boldsymbol{\theta}_0)$ and then apply diffeomorphisms $T_1, T_2 \ldots T_S$

successively so that $\boldsymbol{\theta}_S = T_S \circ T_{S-1} \ldots T_1(\boldsymbol{\theta}_0)$. The transformations $(T_s)_{s=1}^S$, parameterized by ϕ, induce a flexible variational family $Q = \{q_\phi(\boldsymbol{\theta}) | \phi \in \varphi\}$. In this case, the symbol φ denotes the space of parameters for the transformations $(T_s)_{s=1}^S$. We have the following useful relations:

$$\ln q_\phi(\boldsymbol{\theta}_S) = \ln q_0(\boldsymbol{\theta}_0) - \sum_{s=1}^{S} \ln \left| \det \left(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}} \right) \right| \tag{2.3}$$

$$\mathbb{E}_{q_\phi(\boldsymbol{\theta})} h(\boldsymbol{\theta}) = \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} h(T_S \circ T_{S-1} \ldots T_1(\boldsymbol{\theta}_0)) \tag{2.4}$$

(2.3) follows from the change of variable formula and (2.4) is a well known property of expectation (sometimes termed the law of the unconscious statistician). We simplify the maximization of the ELBO in (2.2) as follows:

$$\begin{aligned}
&\max_{q_\phi \in Q} \mathbb{E}_{q_\phi(\boldsymbol{\theta})} \left[\ln p(D|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) - \ln q_\phi(\boldsymbol{\theta}) \right] \\
&= \max_{q_\phi \in Q} \left\{ \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\ln p(D|\boldsymbol{\theta}_S) \pi(\boldsymbol{\theta}_S) \right] + \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\sum_{s=1}^{S} \ln \left| \det \left(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}} \right) \right| \right] - \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\ln q_0(\boldsymbol{\theta}_0) \right] \right\} \qquad (2.5) \\
&= \max_{q_\phi \in Q} \left\{ \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\ln p(D|\boldsymbol{\theta}_S) \pi(\boldsymbol{\theta}_S) \right] + \mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\sum_{s=1}^{S} \ln \left| \det \left(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}} \right) \right| \right] \right\} \qquad (2.6)
\end{aligned}$$

Equations (2.3) and (2.4) jointly imply (2.5). We are essentially re-parametrizing the expectation in terms of the base distribution q_0. In (2.6), we are able to drop $\mathbb{E}_{q_0(\boldsymbol{\theta}_0)} \left[\ln q_0(\boldsymbol{\theta}_0) \right]$ because it is free of the parameter ϕ. In practice, optimizing over $q_\phi \in Q$ effectively becomes optimizing over the parameters ϕ of transformations $(T_s)_{s=1}^S$. We will sometimes refer to ϕ as *flow parameters*. In general, for p-dimensional latent variables $\boldsymbol{\theta}$, calculating the determinant of Jacobian $J_{T_s}(\boldsymbol{\theta}_{s-1}) = \det(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}})$ takes $O(p^3)$ time [28]. Therefore, in addition to $T_1, T_2, \ldots, T_s$ being diffeomorphisms, they are often selected such that computational complexity of calculating $J_{T_s}(\boldsymbol{\theta}_{s-1})$ is $O(p)$.

There are myriad ways in which we can choose the Normalizing Flow transformations. Intuitively, if we choose T_s to be deep neural networks we should be able to approximate almost any well behaved function. But how do we ensure computational feasibility? Neural Auto-regressive Flows (NAF) were proposed in [16] as an attempt at achieving this balance between expressivity and computational feasibility. NAF satisfy the "Universal approximation property". This means

that they can approximate ***any*** probability distribution within an arbitrarily small error margin in the weak convergence sense, provided the width of the neural networks transformations used in the flow are large enough. Further, the auto-regressive structure of these flows ensures the Jacobian determinants can be computed in $O(p)$ time. Note that this is just one among many families of Normalizing Flows. Given these properties we choose to use NAF for our examples in Section 2.6.

2.5 Neural Auto-regressive Flows

Auto-regressive flows are among the most popular Normalizing Flows discussed in the literature. We discuss some of the principals behind auto-regressive Normalizing Flows. We concentrate on describing NAF since we use these for the examples in which we contrast Normalizing Flows aided VI, MCMC and MF-VI.

Continuing with the similar notation, we denote the input from the base distribution by $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \dots \theta_p^0)$ and transformed latent variable by $\boldsymbol{\theta}_1 = (\theta_1^1, \theta_2^1, \dots \theta_p^1)$. For any vector $\boldsymbol{\theta}_s = (\theta_1^s, \theta_2^s, \dots \theta_p^s) \in \mathbb{R}^p$, we let $\boldsymbol{\theta}_{i:j}^s = (\theta_i^s, \theta_{i+1}^s, \dots, \theta_j^s)$ be the sub-vector of $\boldsymbol{\theta}_s$ running from the i^{th} to j^{th} element, where $1 \leq i \leq j \leq p$. Auto-regressive flows are constructed such that each transformed variable $\theta_i^1, 1 \leq i \leq p$ is dependent only on the first i elements $\boldsymbol{\theta}_{1:i}^1$ of $\boldsymbol{\theta}_1$. More specifically, the transformer $T = (\tau_1, \tau_2, \dots \tau_p)$ is made up of p many diffeomorphisms such that:

$$\theta_1^1 = \tau_1(\theta_1^0; c_1)$$
$$\theta_i^1 = \tau_i(\theta_i^0; c_i(\boldsymbol{\theta}_{1:i-1}^0)) \quad 2 \leq i \leq p$$

τ_i is parameterized by the vector $c_i(\boldsymbol{\theta}_{1:i-1}^0)$. The functions $c_i : \mathbb{R}^{i-1} \to \mathbb{R}^m$, $2 \leq i \leq p$ are referred to as conditioners and they enforce the auto-regressive property for the Normalizing Flow. See Figure 2.1 for a visualization of auto-regressive flows. As the name suggests, NAF uses a neural network for τ_i. The 2 types of transformations used are:

(i) **Deep Sigmoidal Flow (DSF)** - This neural network uses a single hidden layer.

(ii) **Dense Deep Sigmoidal Flow (DDSF)** - This uses a deep neural network.

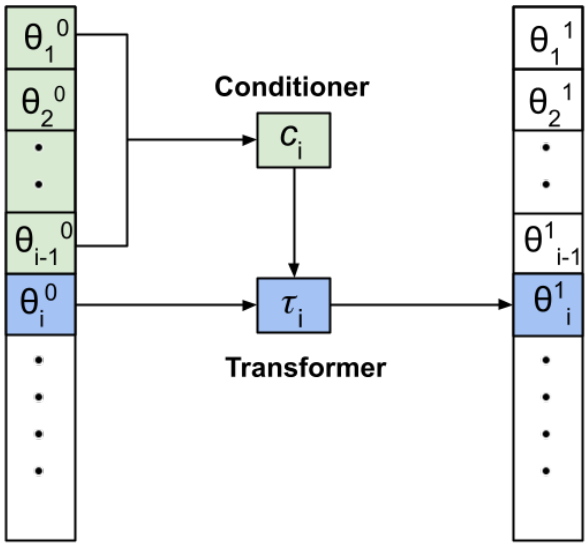


Figure 2.1 *Visualization of auto-regressive flows.*

For readers who are unfamiliar with the topic, think of a neural network as a somewhat complex function that takes some inputs and applies a series of operations and transformations to them. They generally involve multiplication of inputs with weight matrices, translation and and the application of certain "activation" functions. The DSF network is formally defined as:

$$\theta_i^1 = \sigma^{-1}(\boldsymbol{w}_i^\top \sigma(\boldsymbol{a}_i.\theta_i^0 + \boldsymbol{b}_i)) \quad \boldsymbol{a}_i, \boldsymbol{w}_i, \boldsymbol{b}_i \in \mathbb{R}^H \quad 1 \leq i \leq p$$

Here H is the number of nodes in the hidden layer and $\sigma(x) = 1/(1 + e^{-x})$ is an activation function (sigmoid activation). The parameters $\boldsymbol{w}_i,\ \boldsymbol{a}_i,\ \boldsymbol{b}_i$ are the outputs of conditioner networks $c_\phi^{\boldsymbol{w}}(.),\ c_\phi^{\boldsymbol{b}}(.),\ c_\phi^{\boldsymbol{a}}(.) : \mathbb{R}^p \rightarrow \mathbb{R}^{pH}$. Further, $\boldsymbol{a}_i$ and $\boldsymbol{w}_i$ are constrained as $a_{i,j} > 0\ \forall i, j, 0 < w_{i,j} < 1, \sum_j w_{i,j} = 1$. This ensures invertibility of τ_i [16]. Since the DDSF transformation leverages a deep neural network, it has the capacity to be more expressive than DSF, albeit at an increased computational cost.

Until now we have discussed the choice of the transformers τ_i for NAF. To construct the conditioner there are no constraints such as invertibility on the functions c_i. A natural choice is to use a neural network for the conditioner as well. However, using a distinct neural network for each c_i is computationally infeasible as the dimensionality p increases. This is because we have to store and optimize over $p - 1$ networks each with different parameters. [28] discusses a range of conditioners that leverage parameter sharing across c_i. Following [16], we adopt the

popular Masked Conditioner approach. Masked conditioners take $\boldsymbol{\theta}^0_{1:p}$ as inputs to a neural network and calculate all the parameters $c_1, \left(c_i(\boldsymbol{\theta}^0_{1:i-1})\right)_{i=2}^{p}$ for the transformers in a single forward pass. For a network with a single hidden layer, the auto-regressive dependency structure is enforced by multiplying the weight matrices $\mathcal{W}_1$ & $\mathcal{W}_2$ by masking matrices $\mathcal{M}_1, \mathcal{M}_2$ of the same dimension. $\mathcal{M}_1, \mathcal{M}_2$ consist of binary $1-0$ entries such that a 0 entry in $\mathcal{M}_i$ implies the corresponding weighted connection is dropped from the network. Therefore entries in $\mathcal{M}_1, \mathcal{M}_2$ are chosen such that there is no connection between the i^{th} input θ^0_i and $1, 2 \ldots, (mi)$ outputs of the network. Here, m is a multiplier which tells us how many parameters are required for each τ_i. The weight matrices $\mathcal{W}_1 \in \mathbb{R}^{p \times H}$ and $\mathcal{W}_2 \in \mathbb{R}^{H \times mp}$ correspond to the hidden and output layer respectively for the conditioner network. See section 3.4.2 for further technical details on masked auto-regressive conditioners.

2.5.1 Implementation

Recall that for Normalizing **F**lows **A**ided **V**ariational **I**nference (FAVI) we maximize the ELBO (or equivalently minimize the negative of the ELBO):

$$\mathcal{L}(q_\phi) = \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\left[\ln p(D, \boldsymbol{\theta}_S)\right] + \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\left[\sum_{s=1}^{S} \ln \left|\det\left(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}}\right)\right|\right]$$

$(T_s)_{s=1}^S$ and $\boldsymbol{\theta}_S$ depend on ϕ in the equation above. In general, $\mathcal{L}(q_\phi)$ will not have a closed form expression. Additionally, standard co-ordinate wise gradient-ascent algorithms are computationally inefficient for large datasets. As a result, the Stochastic Gradient Ascent (SGA) algorithm is often used for optimizing the ELBO.

SGA is an iterative method that uses the following update for the flow parameters ϕ at step t:

$$\phi_{t+1} = \phi_t + \alpha_t l(\phi_t)$$

The term $l(\phi)$ is a realization for an unbiased estimator of $\nabla_\phi \mathcal{L}(q_\phi)$, the gradient for the ELBO. We can calculate it by sampling $\boldsymbol{\theta}_0^{(1)}, \boldsymbol{\theta}_0^{(2)}, \ldots \boldsymbol{\theta}_0^{(L)}$ from $q_0(.)$ and passing them through the transformations $T_1, T_2, \ldots, T_S$ to get $\boldsymbol{\theta}_S^{(1)}, \boldsymbol{\theta}_S^{(2)}, \ldots, \boldsymbol{\theta}_S^{(L)}$.

$$l(\phi) = \frac{1}{L}\sum_{l=1}^{L}\left[\nabla_\phi \ln p(D, \boldsymbol{\theta}_S^{(l)}) + \sum_{s=1}^{S} \nabla_\phi \ln \left|\det\left(\frac{\partial T_s}{\partial \boldsymbol{\theta}_{s-1}^{(l)}}\right)\right|\right]$$

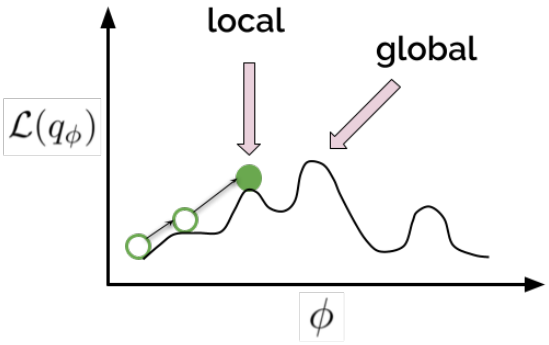


Figure 2.2 *Visualization of the gradient ascent algorithm.*

SGA almost surely converges to a local minimum for non-convex functions and global minimum for pseudo-convex functions when learning rates satisfy $\sum_{t=1}^{\infty} \alpha_t = \infty$ & $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$. ([34], [6]).

In practice, choosing the learning rate α_t is non-trivial. When α_t is too large we may overshoot the maxima and when α_t is too small then SGA will learn too slowly. We use the Adam algorithm [19] which uses an adaptive learning rate that incorporates information about the scale of different components in the parameter vector ϕ. We use a standard normal distribution for $q_0(\boldsymbol{\theta}_0)$.

Note that, the outputs of the Normalizing Flow transformations $\boldsymbol{\theta}_s$ are unconstrained, i.e they belong to $\mathbb{R}^p$, since they are outputs of a neural network. Sometimes the latent variable space is restricted, for example, our model may have a variance parameter $\sigma^2 > 0$. More formally, when $\theta_i \in \mathcal{S} \subset \mathbb{R}$, we apply a final transformation T_{S+1} to constrain $\boldsymbol{\theta}_i^S$. For instance, if $\theta_i > 0$ then we set $\theta_i^{S+1} = \ln(1 + e^{\theta_i^S})$.

2.6 Illustrative Examples

Here, we implement the FAVI algorithm on some examples. We also provide comparisons to MCMC and Mean-Field VI (MF-VI) where applicable. Note that MCMC comprises a wide class of algorithms ranging from the more basic Random-Walk Metropolis Hastings (RW-MH) and Gibbs sampling methods to approaches that make use of gradient information such as the Hamiltonian Monte-Carlo (HMC) [30]. We use either the RW-MH or Gibbs sampling methods as a baseline since these are widely used in classical applications of Bayesian inference. See Section 2.7 for a detailed discussion of contemporary MCMC literature. Section 2.6.1 discusses FAVI

for exponential family models, followed by 2.6.2 in which we sample from un-normalized energy density functions. We then move onto Bayesian linear and logistic regression in 2.6.3 and 2.6.4 respectively. Through these examples we hope to elucidate how FAVI works in different contexts. Code reproducing the results in this chapter can be found at Normalizing-Flows-Review.

2.6.1 The Exponential Family

In many applications of Bayesian inference the complete conditionals $p(\theta_i|\theta_{-i}, D)\ 1 \leq i \leq p$ of latent variables belong to the exponential family $\mathcal{P} = \left\{\frac{h(\theta_i)}{A(\eta)} \exp(\eta^t t(\theta_i))\right\}$. This class of models is known as conditionally conjugate exponential family models and its broad applicability makes it of interest to statistical practitioners. [4] discusses the derivation for Co-ordinate Ascent Variational Inference (CAVI), an MF-VI method, for this class of models. It is natural to extend this discussion to the FAVI algorithm for exponential family models.

FAVI performing well on low-dimensional examples is a necessary but not sufficient condition for them to be reliable for high dimensional VI problems. This motivates our choice of examples:

(i) $(y_i)_{i=1}^n|w \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(w) \quad \pi(w): \ w \sim \text{Beta}(a, b)$

(ii) $(y_i)_{i=1}^n|\mu, \sigma^2 \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2), \pi(\mu, \sigma^2): \ \mu \sim N(0, \tau^2) \ \perp\!\!\!\perp \sigma^2 \sim \text{Inv-Gamma}(v_1, v_2)$

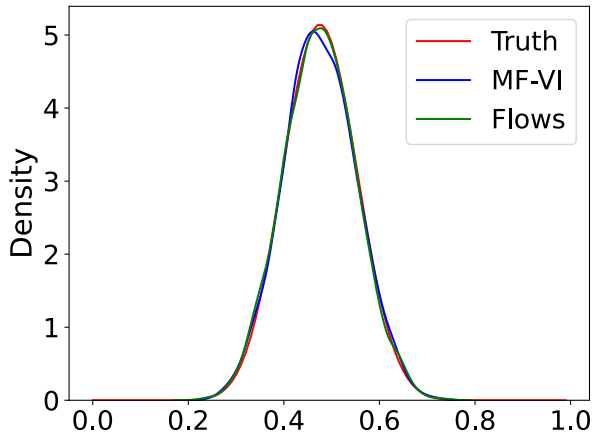
In the first case the posterior has a closed form with which we can compare the density obtained by flows:

$$w|\boldsymbol{y} \sim \text{Beta}(a + n\bar{y}, b + (n - n\bar{y}))$$

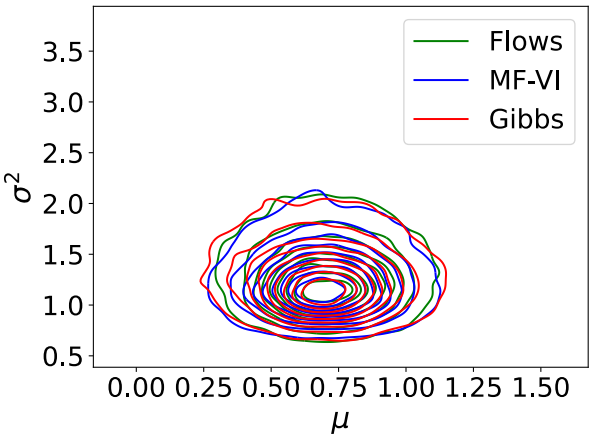
Above we have used $\boldsymbol{y} = (y_1, y_2, \ldots y_n)$ and $\overline{y} = \frac{\sum_{i=1}^n y_i}{n}$.

For the second example, we compare with results obtained by Gibbs sampling. We also include results from MF-VI.

We see from Figure 2.3 that FAVI, MF-VI and Gibbs sampling produce similar results. Figure 2.4 is a demonstration of how the density obtained from FAVI converges to the true distribution over the training epochs. From the plot of ELBO against epochs we see that at around the twentieth epoch there is a plateau. This indicates the density from FAVI has changed shape and is approaching the true distribution.

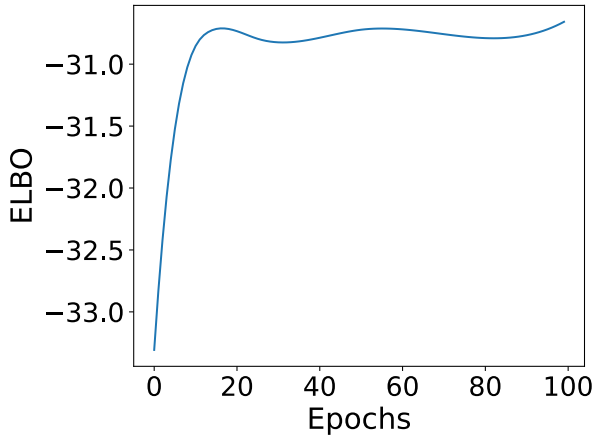


(a) $\Pi(w|\boldsymbol{y})$ - *Bernoulli, Beta*

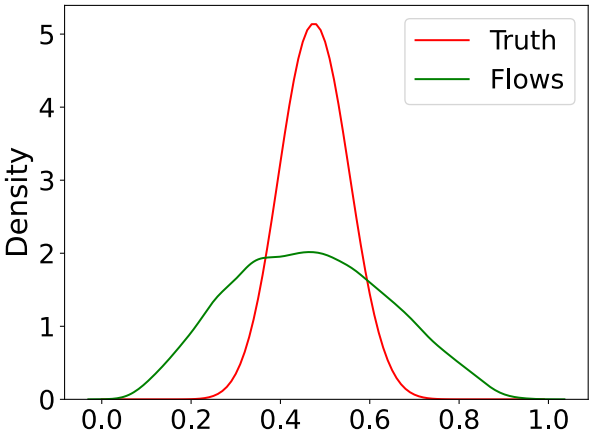


(b) $\Pi(\mu, \sigma^2|\boldsymbol{y})$ - *Normal, Inverse Gamma.*

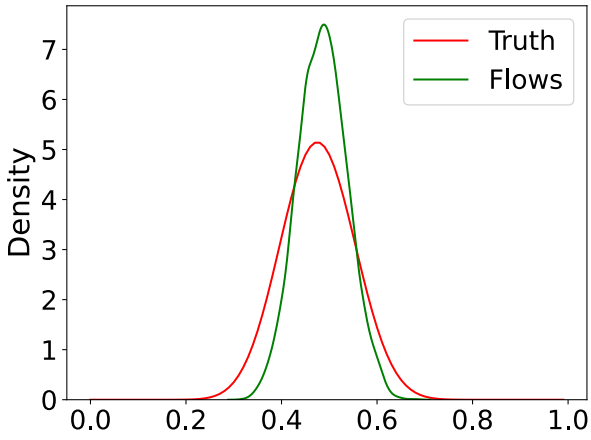
Figure 2.3 *Exponential Family - Posterior Density Plots.*



(a) *ELBO vs Training Epochs.*



(b) $\Pi(w|\boldsymbol{y})$: *epochs* $= 2$.



(c) $\Pi(w|\boldsymbol{y})$: *epochs* $= 20$.

Figure 2.4 *Convergence of FAVI - Bernoulli, Beta.*

2.6.2 Sampling from Multimodal Densities

The primary advantage of Normalizing Flows is their ability to recover highly multimodal target distributions with complex dependencies. In [16], the authors use NAF to sample from multimodal energy density functions, for which the normalizing constant is unknown. They do not however, provide a comparison to MCMC methods. Given that MCMC methods are theoretically guaranteed to converge to the target distribution of interest, we believe it would be useful to include this comparison. We contrast both accuracy and computational time for NAF and the RW-MH Algorithm, for sampling from the energy density functions $U1 - U9$ (see Figure 2.5c).

We compare the 2 methods based on run-time and kernel density estimates. Run-time is measured from the first iteration for the algorithm till convergence. For comparing the densities generated by both methods we calculate kernel density estimates (k.d.e) from the samples. We use a Gaussian kernel, on a grid of size 200×200. We then calculate Square root of Sum of Squared Error ($\sqrt{\mathbf{SSE}}$) for k.d.e over the grid as $\sqrt{\sum_{i=1}^{N=40,000}(\hat{f}(\boldsymbol{x}_i) - f_{\text{True}}(\boldsymbol{x}_i))^2}$. Here, $\hat{f}(.)$ is the kernel density estimator obtained from either the FAVI/RW-MH algorithm and $f_{\text{True}}(.)$ is the true density.[2] This is equivalent to Frobenius norm of errors between true and estimated density on our grid. Results are reported as an average ± standard deviation across 5 different random seeds. We do this to contrast the stability of FAVI (an optimization algorithm), with MCMC - a sampling algorithm.

2.6.2.1 Determining Convergence

For assessing convergence of the RW-MH algorithm we visually inspect the auto-correlation and trace plots. The plots for many energy functions display non-negligible auto-correlation upto lag 40, therefore we thin the samples by 40 and run the chain for 400,000 samples. We choose this run time in order to obtain a sufficient sample size of 10k to get richer kernel density plots. We run FAVI for 15k epochs based on stabilization of the loss function and also generate 10k samples after training. For both the RW-MH and FAVI algorithms there is a degree of subjectivity to determining convergence since we use visual inspection. Empirical convergence criteria such as trace plots, $\hat{R}$ [10] and zero auto-correlation in the samples does not guarantee convergence of the Markov Chain.

[2]Since we do not have the closed form for the true density we normalize the energy functions using numerical integration from SciPy's integrate module.

Although satisfying these criteria is not sufficient for convergence, it is necessary for us to gain confidence that the Markov Chain is approaching the stationary distribution.

2.6.2.2 Results

Table 2.1 reports the average $\sqrt{\text{SSE}}$ of k.d.e for both FAVI and RW-MH algorithms across the best 3 of 5 trials (based on loss) ± standard deviation. We choose best 3 because the loss does not converge in all cases for the FAVI algorithm, due to sensitivity to choice of initialization and the presence of local minima.[3] We observe that the RW-MH algorithm outperforms FAVI in terms of k.d.e metrics by a small-medium margin with one exception, $U6$. We also see from the standard deviations that the RW-MH algorithm is more stable than FAVI, which is highly sensitive to the initialization of flow parameters ϕ. For instance, $U5$ has a standard deviation of $\sqrt{\text{SSE}}$ 0.95 for FAVI and only 0.01 for RW-MH. In terms of computational time, approximately half of the energy functions have run-time of a similar order for the FAVI and RW-MH algorithms. For the remaining functions we observe:

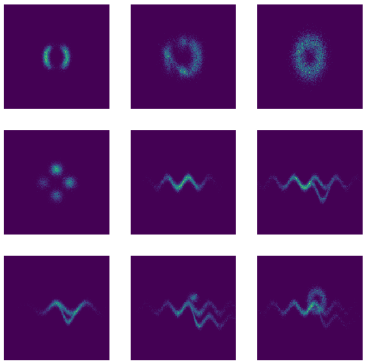
(i) For $U3$ and $U5$ the RW-MH algorithm takes approximately 60% of the run-time that FAVI does.

(ii) For $U4$ and $U9$ the trend is reversed and FAVI takes only 30% of the RW-MH algorithm run-time.

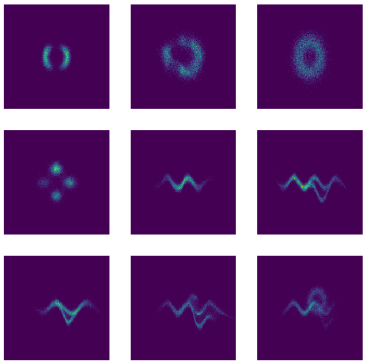
Upon closer examination of the function forms we see that $U3$ and $U5$ are relatively simple functions to evaluate over a particular sample whereas $U4$ and $U9$ are complex functions, being a mixture of multiple densities.[4] $U4$ is a mixture gaussian density with 4 components and $U9$ is a mixture of $U3$ and part of $U8$. Thus, unlike FAVI, the RW-MH algorithm does not scale efficiently as complexity of the target density increases.

[3] We report results across different initializations to contrast the stability across FAVI and MCMC. Further, run-time varies across trials and averaging gives us a better idea of the true run-time.

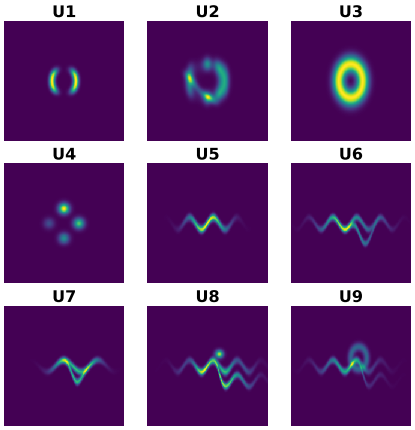
[4] See Appendix A for additional information on these functions.



(a) *Histogram of samples - RW-MH.*



(b) *Histogram of samples - FAVI.*



(c) *True Density.*

Figure 2.5 *Plot of Energy Density Functions* $U1 - U9$.

	Avg $\sqrt{\text{SSE}}$		**Avg Runtime**	
Ef	**FAVI**	**RW-MH**	**FAVI**	**RW-MH**
$U1$	0.98 ± 0.05	0.94 ± 0.01	114 ± 1	119 ± 4
$U2$	0.46 ± 0.03	0.42 ± 0.01	118 ± 7	173 ± 31
$U3$	0.20 ± 0.01	0.18 ± 0.01	109 ± 3	68 ± 8
$U4$	0.62 ± 0.04	0.56 ± 0.02	**122 ± 1**	**465 ± 84**
$U5$	1.72 ± 0.95	1.17 ± 0.01	111 ± 2	69 ± 4
$U6$	**1.21 ± 0.03**	**1.25 ± 0.00**	145 ± 15	212 ± 40
$U7$	1.07 ± 0.01	1.07 ± 0.02	122 ± 7	166 ± 8
$U8$	1.11 ± 0.04	1.10 ± 0.00	127 ± 12	251 ± 13
$U9$	1.25 ± 0.04	1.15 ± 0.01	**131 ± 2**	**303 ± 45**

Table 2.1 *Avg* $\sqrt{SSE} \pm$ *s.d. of k.d.e. for* $U1 - U9$ *(smaller values are better)* | *Avg algorithm runtime in seconds* $\pm$*s.d. for* $U1 - U9$ *(smaller values are better).*

2.6.3 Linear Regression

In this section we implement the FAVI algorithm on a Bayesian Linear Regression example to sample from the posterior of regression parameters given the data, $\Pi(\beta, \sigma^2|D)$. We use the framework below:

$$y_i \sim \boldsymbol{x}_i^\top \beta + \varepsilon_i, \quad \beta \in \mathbb{R}^p, \ \varepsilon_i \overset{\text{i.i.d}}{\sim} N(0, \sigma^2) \ 1 \leq i \leq n$$

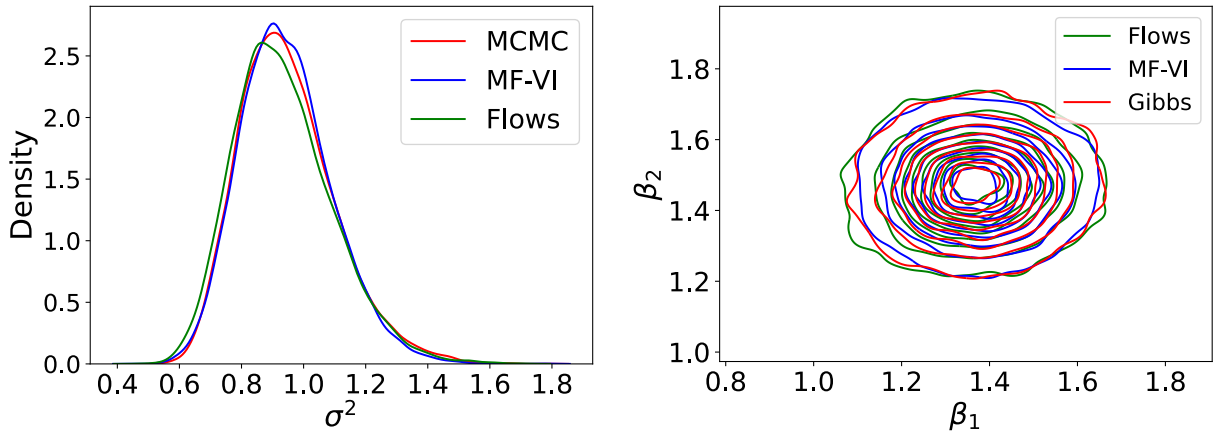
$$\pi(\beta, \sigma^2) : \ \beta \sim N(0, \tau^2 I_p) \ \perp\!\!\!\perp \ \sigma^2 \sim \text{Inv-Gamma}(a, b)$$

As before, we use $\boldsymbol{y} = (y_1, y_2, \ldots y_n)$ and in this case the data $D = (\boldsymbol{y}, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_n)$. We compare FAVI to both MF-VI and the Gibbs Sampling algorithm. We use Gibbs Sampling because it relies on the complete conditionals for latent variables $\pi(\beta|D, \sigma^2)$ and $\pi(\sigma^2|D, \beta)$ which are easily available in this case. Through this example, we can gain some insight on the scalability and accuracy contrast between the 3 methods in a classical statistical set-up. To assess effect of both sample size and dimensionality on the performance of these methods we use a grid of (n, p) combinations. We allow n (sample size) to take values 50, 100 & 200, while p (β dimension) takes values 2, 20, 50, 100.

2.6.3.1 Simulation Details

For our experiments, we simulate the true data generating β_0 from the Uniform$(\frac{1}{2}, 2)$ distribution. We assume $\sigma_0 = \tau = 1$ where σ_0 is the true value of model parameter σ. The p-dimensional predictor variables $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n \in \mathbb{R}^p$ are simulated from a multivariate normal distribution $N(0, \Sigma)$. The covariance matrix $\Sigma = (1 - \rho) I_p + \rho \mathbf{1}\mathbf{1}^\top$, where $\mathbf{1} = (1, 1, \ldots 1)^\top \in \mathbb{R}^p$. The parameter ρ allows us to characterize correlation between predictor variables. For the experiments in this chapter we set $\rho = 0$, since we are primarily interested in the effect of (n, p) combinations on model performance. In chapter 3 we will consider the effect of increasing ρ on the performance of these algorithms. For the cases where $p \geq 20$, we set only 20% of the variables to be non-zero in order to ensure the latent variable space is sparse, that is, $M << p$ where M is the number of non-zero components in β_0.

Similar to Section 2.6.2, convergence of the Gibbs sampling algorithm is determined by a



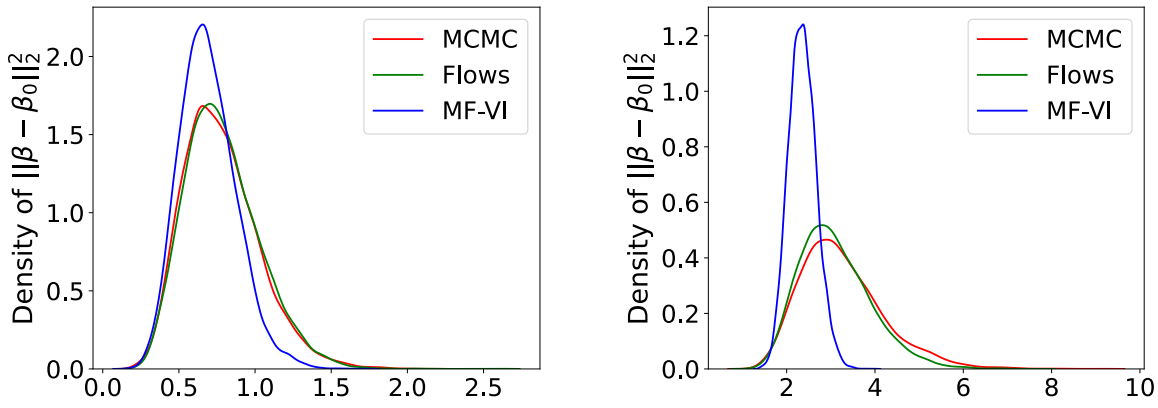
(a) $\Pi(\sigma^2|D)$ (b) $\Pi(\beta_1, \beta_2|D)$

Figure 2.6 ***Linear Regression:*** $n = 100,\ p = 2$

combination of trace and auto-correlation plots. We thin the samples by a factor of 10 to ensure 0 auto-correlation.We initialize β with its O.L.S estimate for faster convergence. Convergence of FAVI and MF-VI is ascertained via stabilization of the loss function.

2.6.3.2 Results

In order to visualize the difference between densities approximated by the 3 approaches (FAVI, MF-VI and Gibbs) we use kernel density plots. For the case where $p = 2$, we can easily visualize the posterior distributions of β and σ^2. For higher dimensional examples we use the kernel density plots for SSE of β; $g(\beta) = \|\beta - \beta_0\|_2^2$ where β is sampled from the posterior $\Pi(\beta|D)$. We present density plots for $n = 100$ and varying p in Figure 2.7. We report the model predictive Root Mean Squared Error ($\sqrt{\textbf{MSE}}$) on test data $\sqrt{\sum_{i=1}^{n_{\text{test}}}(y_i - \hat{y}_i)^2/n_{\text{test}}}$. Here $\hat{y}_i = \boldsymbol{x}_i^\top \hat{\beta}$ is the predicted value for the i^{th} sample based on mean of the posterior samples $\hat{\beta} = 1/N \sum_{n=1}^{N} \beta_n$. The symbol N denotes the number of β samples generated and is set to be 10k. To get a sense of variance of the posterior distribution for β we also report $\overline{s_\beta}$. This is obtained by first computing sample standard deviation from posterior samples for each $\Pi(\beta_i|D)$ as $s_{\beta_i} = (1/(N-1)) \sum_{n=1}^{N}(\beta_i^n - \overline{\beta_i})^2$ for $1 \le i \le p$. We then aggregate these by averaging across dimensions as $\overline{s_\beta} = (1/p) \sum_{i=1}^{p} s_{\beta_i}$. By reporting $\sqrt{\text{MSE}}$ and $\overline{s_\beta}$ we are able to measure both model predictive performance and uncertainty quantification for the 3 methods.



(a) $n = 100$, $p = 20$ (b) $n = 100$, $p = 50$

Figure 2.7 ***Linear Regression:*** *Density Plots of* $||\beta - \beta_0||_2^2$ *where* β *is sampled from* $\Pi(\beta|D)$.

We observe from density plots that for cases where $p = 2$ the difference across 3 algorithms is insubstantial (Figure 2.6). As dimension p increases we see that the FAVI results are reasonably close to Gibbs sampling (the gold standard) but MF-VI produces a peaked distribution, indicating it under estimates uncertainty in the samples (Figure 2.7). Performance metrics in Table 2.2 show that all 3 methods have similar predictive performance. For uncertainty quantification, $\overline{s_\beta}$ shows that MF-VI has lower posterior variance in general than the other two methods, while FAVI is comparable to Gibbs. There are 2 notable exceptions when $n = 100$, $p = 100$ and $n = 50$, $p = 100$. In these cases MF-VI has larger $\overline{s_\beta}$ by 0.01 points than FAVI but this difference is insubstantial.

For the 3 cases where $p \geq n$, the Gibbs sampling algorithm breaks down due to the instability of the following matrix inversion: $(I_p/\tau^2 + X^\top X/\sigma^2)^{-1}$. From error metrics we see that both these algorithms have somewhat poor predictive performance for these cases. It is to be expected that a naive model such as this (that does not incorporate sophisticated regularization) may perform rather poorly in high dimension. A more interesting observation is that since FAVI, like MF-VI, does not require the matrix inversion step and can specify an arbitrarily flexible family of densities, it could be a promising alternative to Gibbs sampling in such a set-up.[5]

Table 2.2 reports average algorithm run-times ± s.d. across 5 trials. In general, MF-VI runs the fastest, followed by Gibbs sampling and then FAVI. However, for the highlighted cases of $p = 50$

[5]There are modifications of Gibbs sampling that can circumvent this and a more thorough exploration is required.

	Pred $\sqrt{\text{MSE}}$			Avg. β s.d. $\overline{s_\beta}$		
n, p	**Gibbs**	**FAVI**	**MF-VI**	**Gibbs**	**FAVI**	**MF-VI**
$(50, 2)$	1.06	1.06	1.06	0.17	0.17	0.16
$(50, 20)$	0.88	0.92	0.88	0.24	0.23	0.17
$(50, 50)$	*	3.06	3.41	*	0.47	0.46
$(50, 100)$	*	2.93	2.64	*	0.77	0.78
$(100, 2)$	1.04	1.04	1.04	0.11	0.11	0.11
$(100, 20)$	1.05	1.05	1.05	0.14	0.14	0.12
$(100, 50)$	1.93	1.90	1.93	0.17	0.16	0.11
$(100, 100)$	*	2.64	2.92	*	0.46	0.47
$(200, 2)$	1.07	1.07	1.07	0.08	0.09	0.08
$(200, 20)$	1.10	1.09	1.10	0.09	0.09	0.08
$(200, 50)$	1.26	1.26	1.26	0.10	0.10	0.08
$(200, 100)$	1.86	1.87	1.86	0.14	0.13	0.09

Avg Runtime (in seconds)		
Gibbs	**FAVI**	**MF-VI**
55 ± 0	252 ± 8	22 ± 1
79 ± 5	374 ± 21	23 ± 1
*	606 ± 41	122 ± 10
*	1055 ± 84	922 ± 69
59 ± 4	267 ± 10	24 ± 2
85 ± 5	383 ± 24	24 ± 1
953 ± 54	**549** ± 23	36 ± 2
*	1087 ± 56	778 ± 80
57 ± 3	254 ± 13	23 ± 1
80 ± 0	568 ± 10	23 ± 0
997 ± 110	**567** ± 31	24 ± 2
833 ± 37	1011 ± 28	35 ± 1

Table 2.2 ***Results for Linear Regression. Left:*** *Model Predicted square root of mean square* $\sqrt{MSE}$ *(smaller values are better) | Avg. s.d. of* β *samples.* ***Right:*** *Avg algorithm runtime ± s.d. over 5 trials (smaller values are better). We use * when the result can't be computed.*

and $n \geq 100$ this trend switches and Gibbs sampling becomes slower than FAVI. Given that Gibbs sampling breaks down in high-dimension it is difficult to discern a pattern and contrast scalability of the 3 methods.

2.6.4 Logistic Regression

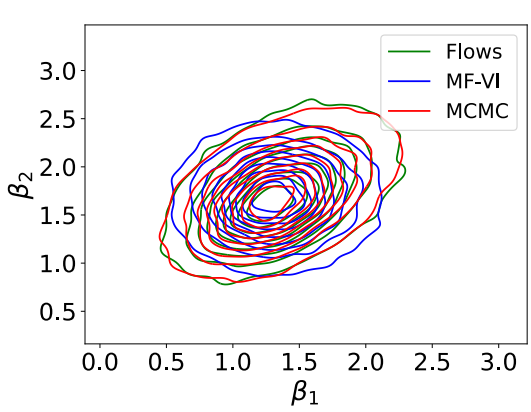
We consider the following model:

$$y_i \sim \text{Bernoulli}(p_i), \quad p_i = \frac{e^{x_i^\top \beta}}{1 + e^{x_i^\top \beta}}$$

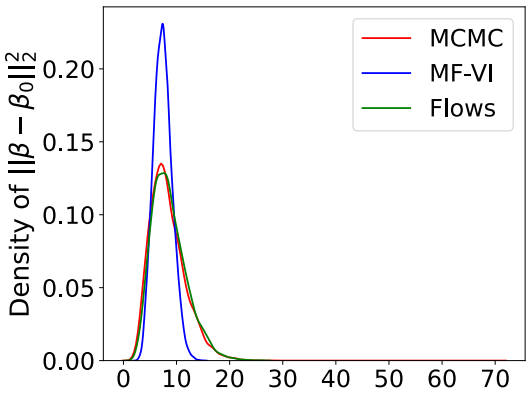
$$\pi(\beta): \quad \beta \sim N(0, \tau^2 I_p)$$

We use the same simulation set-up for β and $\boldsymbol{x}_i$ as of Section 2.6.3.1 on linear regression. Here we use RW-MH instead of Gibbs sampling because we no longer have closed form complete conditionals. Maintaining consistency with previous experiments, we use trace and auto-correlation plots to decide on convergence.[6] We initialize the RW-MH algorithm with the maximum likelihood estimates for β. We use plots and metrics as in 2.6.3, replacing $\sqrt{\text{MSE}}$ by Accuracy.

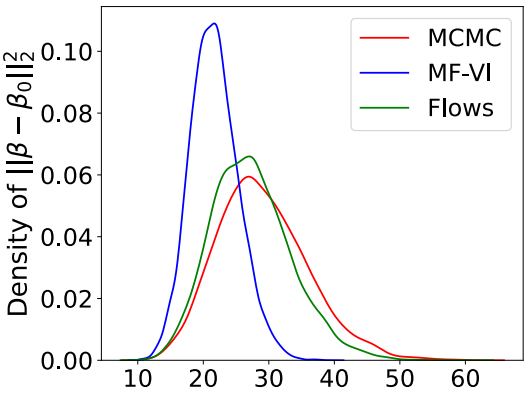
[6]For most cases the auto-correlation after thinning is between 0.0 – 0.2, however when $p = 100$ we allow auto-correlation of 0.4 given computational considerations.



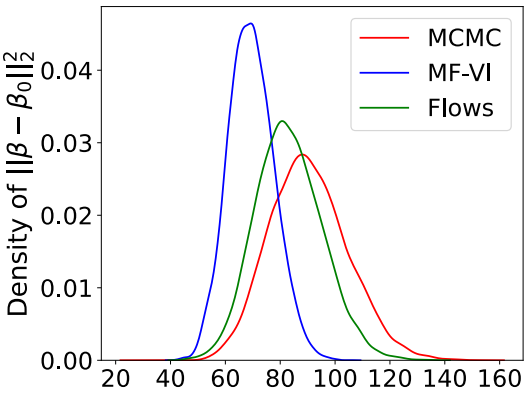
(a) $\Pi(\beta_1, \beta_2|D)$ *when* $n = 100,\ p = 2$.



(b) Density of $\|\beta - \beta_0\|_2^2$, $\beta \sim \Pi(\beta|D)$ when $n = 100$, $p = 20$.



(c) Density of $\|\beta - \beta_0\|_2^2$, $\beta \sim \Pi(\beta|D)$ when $n = 100$, $p = 50$.



(d) Density of $\|\beta - \beta_0\|_2^2$, $\beta \sim \Pi(\beta|D)$ when $n = 100$, $p = 100$.

Figure 2.8 ***Logistic Regression:*** *Density Plots.*

2.6.4.1 Results

Density plots for β when $n = 100, p = 2$ are presented in Figure 2.8a. From the contour plot we see that MF-VI does not capture the elliptical structure of the joint distribution of β which the other 2 methods display. For higher dimensions, kernel density plots of β SSE, $\|\beta - \beta_0\|_2^2$ when $\beta \sim \Pi(\beta|D)$ are presented in Figure 2.8. Similar to the trend displayed by Gaussian linear regression, we see that kernel density plots for MF-VI seem to center on a lower SSE. The FAVI and RW-MH algorithms perform similarly with respect to uncertainty quantification as dimension p increases but MF-VI has lower aggregate posterior variance for β (See Table 2.3). All 3 methods display identical model predictive Accuracy given by $\sum_{i=1}^{n_{\text{test}}} \mathbb{I}\{\hat{y}_i = y_i\}/n_{\text{test}}$.

The plot of average run-time across 5 trials against dimension p presents an interesting contrast

	Accuracy			Avg. β sd. $\overline{s_\beta}$		
n, p	**MH**	**FAVI**	**MF-VI**	**MH**	**FAVI**	**MF-VI**
(50, 2)	0.80	0.80	0.80	0.41	0.40	0.39
(50, 20)	0.90	0.90	0.90	0.56	0.55	0.47
(50, 50)	0.60	0.60	0.60	0.77	0.73	0.61
(50, 100)	0.50	0.50	0.50	0.88	0.85	0.74
(100, 2)	0.60	0.60	0.60	0.36	0.36	0.32
(100, 20)	0.65	0.65	0.65	0.40	0.39	0.34
(100, 50)	0.95	0.95	0.95	0.59	0.55	0.44
(100, 100)	0.70	0.70	0.70	0.77	0.71	0.56
(200, 2)	0.85	0.85	0.85	0.25	0.25	0.25
(200, 20)	0.85	0.85	0.85	0.25	0.25	0.25
(200, 50)	0.55	0.55	0.55	0.40	0.38	0.31
(200, 100)	0.78	0.78	0.78	0.61	0.54	0.42

Table 2.3 ***Logistic Regression:*** *Model Accuracy (larger values are better) | Avg. s.d. of β samples. Here RW-MH is abbreviated as MH.*

(Figure 2.9). As expected, MF-VI scales the best, since it has run-time of approximately the same order regardless of dimension. RW-MH algorithm scales poorly ranging from 40s for $p = 2$ to approximately 1000s when $p = 100$. FAVI on the other hand, only has a run-time 400s when $p = 100$, less than half the time of the RW-MH algorithm (see Appendix B for a detailed table containing run-time comparisons for logistic regression). Thus, FAVI is scalable relative to the RW-MH algorithm and performs much better than MF-VI at approximating densities in their entirety, beyond just measures of central tendency.

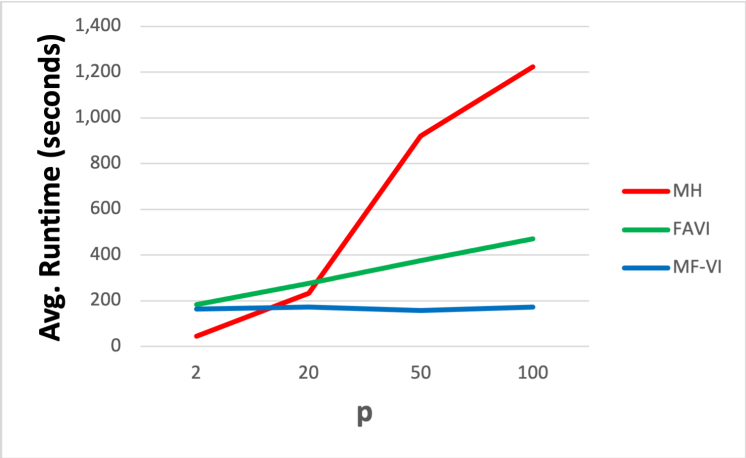


Figure 2.9 ***Logistic regression:*** *Average Runtime vs β dimension when $n = 200$.*

2.7 Looking Ahead

This chapter has discussed how Normalizing Flows are a useful tool in probabilistic modeling. In the rest of this section we will summarize the results of our preliminary experiments and elaborate on some interesting open problems concerning Normalizing Flows and their use for statistical inference. This will motivate the next chapter of this dissertation.

The examples covered in Section 2.6 confirm that FAVI lies somewhere between basic MCMC methods (RW-MH, Gibbs sampling) and MF-VI with respect to accuracy and scalability. They can approximate multimodal densities and come reasonably close to MCMC for uncertainty quantification while retaining some scalability. An exciting feature of FAVI is a high degree of flexibility over the desired levels of expressivity and scalability for the probabilistic model. This is due to our ability to select flow depth, type of transformation, and the number of flow parameters ϕ.

There are still many challenges remaining in this area. There is no rigorous study on the scalability vs. expressivity trade-off in the literature. Many Normalizing Flow families such as NAF have universal approximation properties which allow them to approximate any distribution arbitrarily well, given enough flow depth and complexity. This does not, however, provide answers on the flow depth and complexity required to achieve desired levels of accuracy. In contrast, MCMC methods have effective sample size measures, which indicate the minimum amount of samples needed to obtain the required quality in posterior samples.

Another research direction would be improving FAVI's scalability without a significant accuracy loss. We observe that FAVI can be computationally expensive in Bayesian inference if the likelihood $p(D|\boldsymbol{\theta})$ is difficult to evaluate. There are already efforts in this direction. [43] replaces the likelihood with a surrogate likelihood that we can learn while training the flow transformations. In addition, we see in Section 2.6.2, that FAVI can be highly sensitive to the initialization of flow parameters ϕ. Thus FAVI can take longer to converge to the minima with bad initializations. It follows that a beneficial avenue of research would be going beyond naive initializations for FAVI.

As discussed in Section 2.6, we have used the RW-MH and Gibbs sampling algorithms for our experiments. However, there exist a range of other MCMC algorithms in the literature. Among the most popular is the HMC algorithm [30], and its self tuning variant, the No-U-Turn sampler (NUTS) [14]. These methods have achieved considerable success by leveraging gradient information to make jumps through the state space for the target distributions. In fact, HMC scales at $O(p^{1/4})$ iterations to achieve 2 nearly independent samples in comparison to $O(p)$ time for RW-MH. Thus, HMC is considered to be the gold standard for unimodal high-dimensional regimes, if computationally feasible. [44] contains a comparison of FAVI and HMC on 13 Bayesian linear regression models. Their results indicate that FAVI is competitive with HMC, having a lower MSE in 5 of 13 models. [26] shows that for highly multimodal distributions the above scaling regime need not hold. Specifically, HMC and the RW-MH algorithm behave the same way, with their spectral gaps decaying at the same rate. Thus, FAVI has the potential to compete with HMC for multimodal densities if we have a limited computational budget. A more rigorous, wide scale exploration of how FAVI compares to gradient based MCMC methods is essential.

Normalizing Flows can also be used to aid MCMC sampling. We expand on some existing ideas to do this. For multimodal target distributions, the MCMC chains may converge slowly, and samples are highly correlated.[7] The MH algorithm, uses a "proposal density" $p(\boldsymbol{\theta})$ from which we generate candidates for posterior samples. The proposal density is often selected to be easy to sample from, for example, the gaussian distribution. Unfortunately, when the target density

[7] We see this occur for the mixture Gaussian density $U4$ in section 2.6.2

has a complicated geometry, this results in slow exploration of the state space. To address this, we can use Normalizing Flows to shift the proposal density space to a "distorted" space. This is possible because Normalizing Flows are nothing but a re-shaping of one density into another. The MH algorithm is then able to move faster through this "distorted" space. Recently, Inverse Auto-regressive Flows have been used to aid HMC sampling [15].

Until now, we have discussed how Normalizing Flows can be used for learning continuous distributions. What happens for discrete probability distributions? There is an equivalent change of variable formula for flows on discrete distributions:

$$p_Z(\boldsymbol{z}) = p_U(T^{-1}(\boldsymbol{z}))$$

$T : \mathcal{U} \rightarrow \mathcal{Z}$ is a bijection between two discrete spaces $\mathcal{U}, \mathcal{Z}$. However, some issues exist with using the above for learning discrete distributions. For one, we rely heavily on the base distribution for expressivity in discrete flow models. We need to incorporate dependencies across variables into the base distribution itself. Further, there is no research on modeling joint discrete-continuous distributions. We expect this to be a popular avenue of research in the near future.

Normalizing Flows are a significant advancement for probabilistic modeling, particularly for VI. This area is in the nascent stages, and many issues need to be tackled. In order to enable the wider adoption of Normalizing Flows aided VI for Bayesian inference, it is crucial to understand the statistical properties of the variational posterior obtained from FAVI. This will motivate Chapter 3 of this dissertation.

CHAPTER 3

STATISTICAL PROPERTIES OF THE FAVI POSTERIOR: A CASE STUDY WITH LINEAR REGRESSION

As discussed in the previous chapters, Variational Inference (VI) has emerged as a popular tool to sample from probability distributions for which the normalizing constant is unknown. This problem frequently occurs in applications requiring uncertainty quantification with Bayesian statistics. In order for this chapter to be self-contained, we briefly review some basic concepts of Bayesian statistics and VI below.

In Bayesian statistics, all inference about an unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^p$, governing a random data generating process, comes from the posterior distribution:

$$\Pi(\boldsymbol{\theta}|D) = (p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta}))/(\int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta})$$

The symbol $\pi(\boldsymbol{\theta})$ denotes the prior distribution, $p(D|\boldsymbol{\theta})$ is the likelihood and D is the observed data. The stochasticity in $\boldsymbol{\theta}$ induced by the prior facilitates uncertainty quantification for our estimates.

In applications where $m(D) = \int_{\boldsymbol{\theta}} p(D|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$ is intractable, we resort to approximate inference methods such as VI, wherein the problem of sampling from the target posterior distribution is converted to the following optimization:

$$q_{\phi^*} \in \arg\min_{q_\phi \in Q} KL(q_\phi \,||\, \Pi(.|D)) \tag{3.1}$$

Mean-Field VI (MF-VI) assumes any $q_\phi^{\text{MF}} \in Q^{\text{MF}}$ can be factorized as $q_\phi^{\text{MF}}(\boldsymbol{\theta}) = \prod_{i=1}^p q_{\phi_i}^{\text{MF}}(\theta_i)$, in order to facilitate computational efficiency. Structured VI (SVI) allows for some degree of dependencies among θ_i by estimating a non-diagonal covariance $\Sigma \in \mathbb{R}^{p\times p}$, but is still limited to unimodal distributions.

We recollect that, Normalizing **F**lows **a**ided **VI** (FAVI) generates a family of distributions Q^{NF} by sampling from a base distribution $q_0(\boldsymbol{\theta}_0)$ and applying differentiable and invertible mappings $T_s : \mathbb{R}^p \to \mathbb{R}^p$ such that $\boldsymbol{\theta}_S = T_S \circ T_{S-1} \cdots \circ T_1(\boldsymbol{\theta}_0)$. If $\varphi \subseteq \mathbb{R}^m$ denotes the space of parameters for the transformations $\{T_s\}_{s=1}^S$, then any $\phi \in \varphi$ results in a distribution $q_\phi^{\text{NF}}(\boldsymbol{\theta}_S)$. We then choose an optimal $q_{\phi^*}^{\text{NF}}$ from $Q^{\text{NF}} = \{q_\phi^{\text{NF}} \mid \phi \in \varphi\}$ based on the minimization in (3.1) (see section 2.4.1). As

before, we limit our scope to auto-regressive flows. For explanations of the mathematical notation relating to VI used throughout the remaining chapters of this dissertation, refer to Table 3.1.

FAVI is widely used to address the limitations of MF-VI and SVI, with the landmark paper [33] being cited $\geq$ 3500 times since its 2015 publication. In chapter 2 we mentioned that certain classes of Normalizing Flows (such as neural auto-regressive flows), are universal approximators. This means that they can recover **any** distribution on $\mathbb{R}^p$ arbitrarily well, provided T_s are expressive enough [28]. Despite this, the statistical properties of the variational posterior $q^{\text{NF}}_{\phi^*} \in Q^{\text{NF}}$ are not well understood. The universal approximation results for Normalizing Flows (NFs) do not account for the optimization (3.1). It is thus, theoretically unclear how well $q^{\text{NF}}_{\phi^*}$ emulates $\Pi(\boldsymbol{\theta}|D)$. There is also little work characterizing the gains in exact recovery of the posterior samples from FAVI at varying complexity of T_s, especially with respect to uncertainty quantification.

In this chapter, we provide a first attempt at answering these questions within the specific context of the Bayesian Linear model:

$$\boldsymbol{y} = X\boldsymbol{\theta} + \epsilon, \quad \boldsymbol{\theta} \in \mathbb{R}^p \sim \pi(\boldsymbol{\theta}) \tag{3.2}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is the response vector, $X \in \mathbb{R}^{n\times p}$ is the design matrix of predictor variables, $\boldsymbol{\theta}$ is unknown and $\epsilon \in \mathbb{R}^n \sim N(0, \sigma^2 I)$ is random noise. We choose this set-up due to: (i) its simplicity; (ii) wide applicability; (iii) the surge of interest in using VI for high-dimensional Bayesian regression [31], [7], [29]. This framework also facilitates quantifying the accuracy of the credible interval[1] coverage based on the FAVI posterior at different complexity levels of T_s (See Corollary 3.2.1.1).

3.1 Contribution

For simplicity, we begin by considering the case of 2 predictors, that is, $p = 2$ and $X = [\boldsymbol{x}_1, \boldsymbol{x}_2]$. Section 3.5 considers extensions to the case $p > 2$. Let Q^{IAF} be the family of distributions generated by Inverse Auto-Regressive Flows (IAF), a type of NF introduced in [21]. Let the linear model (3.12), with observed data $D = (X, \boldsymbol{y})$, have target posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$. We derive the variational posterior $q^{\text{IAF}}_{\phi^*} \in Q^{\text{IAF}}$ that approximates $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ to learn how closely it models the true posterior.

[1] Credible intervals are in some sense the Bayesian counterpart to frequentist confidence intervals. This concept is formally defined in Section 3.2.1.1.

Symbol	Description
$KL(q \mid\mid p)$	Kullback-Leibler divergence between 2 distributions.
Q	Variational family of distributions.
q_ϕ	A member of Q with variational parameters ϕ.
$q_{\phi*}$	Optimal variational distribution with parameters ϕ^*.
q_0	Base distribution for a Normalizing Flow (NF).
T	NF transformation applied to samples from q_0.
c_ϕ	Conditioner network for inverse auto-regressive flows (IAF).
$Q^{\{\cdot\}}$	Generating distributional family ($\cdot$ is NF, IAF, MF (mean-field)).
$q_\phi^{\{\cdot\}}$	A member of $Q^{\{\cdot\}}$ ($\cdot$ is NF, IAF, MF).
$q_{\phi^*}^{\{\cdot\}}$	Optimal variational distribution from $Q^{\{\cdot\}}$ ($\cdot$ is NF, IAF, MF).

Table 3.1 Glossary of notation relating to variational inference. More exhaustive information on the basic statistical notation used in this chapter is included in Table 3.2.

Our results are presented at different complexity levels of a single IAF transformation[2] T and provide an initial framework for analyzing the accuracy vs computational cost trade-off for the FAVI posterior. We also derive the loss in uncertainty quantification (credible interval coverage), from using the approximation $q_{\phi^*}^{\text{IAF}}$ in lieu of the true posterior, as a function of the correlation ρ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We contrast the loss in credible interval coverage from FAVI to that of MF-VI (since MF-VI is the most commonly used VI algorithm in the literature). Adaptations to other VI approaches are briefly covered in the Discussion. Although there are simulation studies in [31], [7] showing that MF-VI under-estimates the posterior variance as ρ increases, to our knowledge the impact of this on the credible interval coverage has not yet been mathematically quantified and we are the first to do so.[3]

3.1.1 Assumptions

We assume a $N(0, I)$ prior on $\boldsymbol{\theta}$ and a known σ^2. This implies $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ has a closed form, which helps to obtain analytic results. Specifically, $\boldsymbol{\theta}|X, \boldsymbol{y} \sim N(m_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$ where $m_{\boldsymbol{\theta}}$ and $\Sigma_{\boldsymbol{\theta}}$ are given by:

$$m_{\boldsymbol{\theta}} = \Sigma_{\boldsymbol{\theta}}^{-1} \frac{X^\top \boldsymbol{y}}{\sigma^2}, \quad \Sigma_{\boldsymbol{\theta}} = \left(\frac{X^\top X}{\sigma^2} + I \right)^{-1} \tag{3.3}$$

[2]For the rest of this chapter we assume $S = 1$ and drop the subscript from T_s.

[3]Code reproducing Figure 3.1 and 3.2 subplots can be found at FAVI_for_Bayesian_Regression.

To generate Q^{IAF} we sample $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots \theta_p^0)^\top \sim N(0, I)$ and then apply a transformation $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0)$ as follows:

$$\begin{aligned} &\theta_i = b_i + a_i \theta_i^0, \quad 1 \le i \le p \\ &b_i = c_\phi^b(\boldsymbol{\theta}_{1:i-1}^0) \quad \text{and} \quad a_i = h(c_\phi^a(\boldsymbol{\theta}_{1:i-1}^0)), \quad 2 \le i \le p \end{aligned} \tag{3.4}$$

Here, $\left\{\{c_\phi^b(\boldsymbol{\theta}_{1:i-1}^0), c_\phi^a(\boldsymbol{\theta}_{1:i-1}^0)\}_{i=2}^p, b_1, a_1\right\}$ are the outputs of "conditioner" neural networks $\boldsymbol{c}_\phi^b(.),\ \boldsymbol{c}_\phi^a(.) : \mathbb{R}^p \to \mathbb{R}^p$ that enforce the auto-regressive structure for the NF, wherein θ_i depends only on the sub-vector $\boldsymbol{\theta}_{1:i-1}$ (see Section 3.4.2 for details). The function $h : \mathbb{R} \to \mathbb{R}^+$ is invertible and ensures $a_i > 0$. We use the width (number of hidden nodes) of the networks $\boldsymbol{c}_\phi$ as the measure of complexity of T. Although we have chosen IAF, a very simple class of auto-regressive NFs, we see later that they are sufficient for approximating the Gaussian posterior. We also assume a ReLU function $g(x) = \max(x, 0)$, one of the more often used piecewise polynomial activations for $\boldsymbol{c}_\phi$. Extensions to other activations are left to future work.

3.2 Main Results

Theorem 3.2.1. *Consider the posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ defined in (3.3). Let $q^{IAF}_{\phi^*} \in Q^{IAF}$, the optimal approximate posterior for $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$, be defined as:*

$$q^{IAF}_{\phi^*} \in argmin_{q^{IAF}_{\phi} \in Q^{IAF}} KL(q^{IAF}_{\phi} \,||\, \Pi(.|X, \boldsymbol{y}))$$

where Q^{IAF} is the family generated by IAF with K hidden nodes in the shallow network $\boldsymbol{c}_\phi$ from (3.4). Then we have:

(i) ***Case when $K = 1$:***

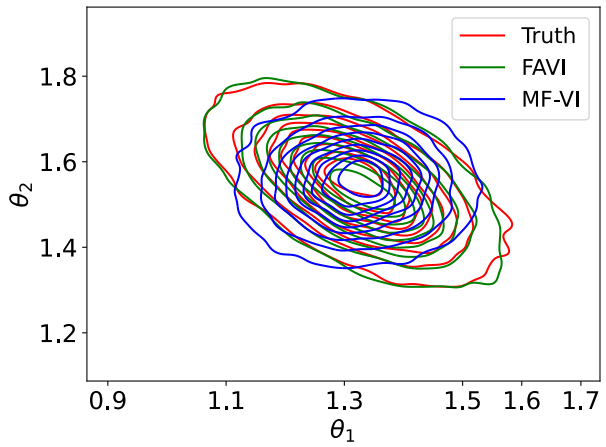
a) $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \neq 0 \implies KL(q^{IAF}_{\phi} \,||\, \Pi(.|X, \boldsymbol{y})) > 0$ *for all* $q^{IAF}_{\phi} \in Q^{IAF}$ *and* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 = 0 \implies KL(q^{IAF}_{\phi^*} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$.

b) $\exists$ *some combination of parameters* $\psi^*(\tilde{t}) = (\phi^*_{-(\tilde{t})}, \tilde{t})$ *s.t* $q^{IAF}_{\psi^*(\tilde{t})}$ *has mean* $m_{\boldsymbol{\theta}}$ *and covariance* $\Sigma_{\psi^*(\tilde{t})}$. *When* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 > 0$ *then* $\lim_{\tilde{t} \to -\infty} \Sigma_{\psi^*(\tilde{t})} = \Sigma_{\boldsymbol{\theta}}$ *element-wise and* $\lim_{\tilde{t} \to -\infty} KL(q^{IAF}_{\psi^*(\tilde{t})} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$. *If* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 < 0$, *then we have the same limits provided* $\tilde{t} \to +\infty$.

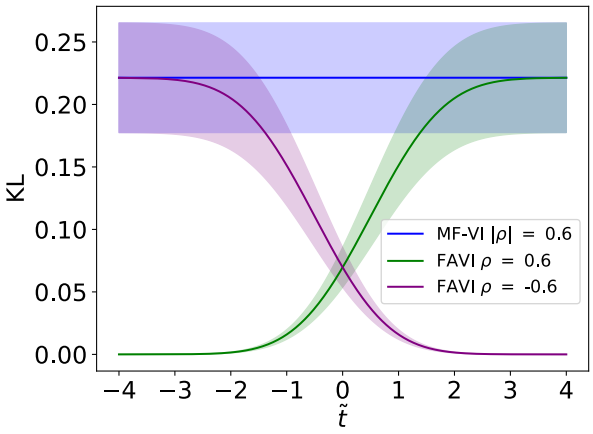
(ii) ***Case when $K = 2$:*** $q^{IAF}_{\phi^*}(\boldsymbol{\theta}) = \Pi(\boldsymbol{\theta}|X, Y) \; \forall \; \boldsymbol{\theta} \in \mathbb{R}^2$ *and* $KL(q^{IAF}_{\phi^*} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$.

Here, $\tilde{t}$ in (ib) is a re-scaled bias for the hidden node in $\boldsymbol{c}^b_\phi$ (see section 3.19) and $\phi^*_{-(\tilde{t})}$ are the remaining IAF parameters.

Since $KL(q||p) = 0 \iff q = p$ almost everywhere (a.e), (ia) tells us that with one hidden node in $\boldsymbol{c}_\phi$ we cannot recover the true posterior entirely unless $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 = 0$. However, (ib) says we can find a sequence of distributions for which the KL approaches 0. Thus, choosing $\tilde{t}$ appropriately ensures we *almost* entirely recover $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ when $K = 1$. The density plot of the true and approximate posteriors in Figure 3.1a shows that MF-VI cannot recover the ellipsoid shape of $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ at correlation $\rho = -0.6$ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, but FAVI with IAF and $K = 1$ tracks the true posterior closely. Figure (3.1b) displays the limiting behaviour of the KL described in (ib). The flip in behavior at $\rho = 0.6$ vs $\rho = -0.6$ in can be explained by the fact that the optimal KL is actually a function of $\tilde{t}$ and $\text{sign}(\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)$ (see proof of Theorem 3.2.1 in 3.4.3.2). In order to



(a) Density plot of samples from $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$, $q^{\text{MF}}_{\phi^*}$ and $q^{\text{IAF}}_{\phi^*}$ when $K = 1$. We set $n = 200$, $\sigma = 1$ and the data generating $\theta_0 \sim U(0.5, 2)$. The rows $\boldsymbol{z}_j$ of X are simulated as $\boldsymbol{z}_j \overset{\text{i.i.d}}{\sim} N(0, \rho\mathbf{1}\mathbf{1}^\top + (1-\rho)I)$, $\rho = -0.6$.



(b) Plot of the mean KL (bold lines) with ±1SD (shaded area) against $\tilde{t}$ for $q^{\text{IAF}}_{\phi^*}$ at $K = 1$ and $q^{\text{MF}}_{\phi^*}$ at $n = 200$. The mean and SD are taken across 100 simulations. Simulation hyper-parameters are the same as Figure 3.1a with $\rho = \pm 0.6$.

Figure 3.1 Visualizing the behaviour of $q^{\text{MF}}_{\phi^*}$ and $q^{\text{IAF}}_{\phi^*}$ when $K = 1$.

completely recover $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ we require $K = 2$. We prove these results by analyzing the form of $q^{\text{IAF}}_{\phi} \in Q^{\text{IAF}}$ and choosing ϕ^* that minimizes the KL. The proof is presented in Section 3.4.3.2.

3.2.1 Uncertainty Quantification

We now move on to the loss in uncertainty quantification when using credible intervals from FAVI or MF-VI in place of the true posterior. A $1-\alpha$ credible interval for θ_i is a range of values $(L^{(i)}_\alpha, U^{(i)}_\alpha)$ such that θ_i belongs to this interval with probability $1-\alpha$ under the posterior distribution. Let q_{ϕ^*} be the variational posterior and $F^{(i)}_{\phi^*}$ denote the marginal cumulative distribution function (cdf) for $q_{i,\phi^*}(\theta_i) = \int_{\boldsymbol{\theta}_{-(i)}} q_{\phi^*}(\boldsymbol{\theta}) d\boldsymbol{\theta}_{-(i)}$. Then the $1-\alpha$ equal-tailed credible interval for θ_i is:

$$(F^{(i)^{-1}}_{\phi^*}(\frac{\alpha}{2}), F^{(i)^{-1}}_{\phi^*}(1-\frac{\alpha}{2})) \tag{3.5}$$

The actual coverage of the $1-\alpha^{(i)}_\Pi$ credible interval in (3.5) under the true posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ is then given by:

$$1-\alpha^{(i)}_\Pi = \Phi\Big(\frac{F^{(i)^{-1}}_{\phi^*}(1-\frac{\alpha}{2}) - m_{\boldsymbol{\theta}i}}{(\Sigma_{\boldsymbol{\theta}})^{\frac{1}{2}}_{ii}}\Big) - \Phi\Big(\frac{F^{(i)^{-1}}_{\phi^*}(\frac{\alpha}{2}) - m_{\boldsymbol{\theta}i}}{(\Sigma_{\boldsymbol{\theta}})^{\frac{1}{2}}_{ii}}\Big) \tag{3.6}$$

where $m_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}$ are as in (3.3) and Φ is the $N(0,1)$ cdf. The difference $(1-\alpha) - (1-\alpha^{(i)}_\Pi)$ denotes the loss in uncertainty quantification when replacing the true posterior with its approximation.

Corollary 3.2.1.1 expresses $1-\alpha_{\Pi}^{(i)}$ as function of the correlation ρ between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, for FAVI and MF-VI.

Corollary 3.2.1.1. *Let $1-\alpha$ be the coverage for the equal-tailed credible intervals for $\theta_i,\ i=1,2$ specified in (3.5), corresponding to the approximate posterior q_{ϕ^*}. Let $1-\alpha_{\Pi}^{(i)}$ defined in (3.6) be the actual coverage of this interval under the true posterior. Let $\boldsymbol{z}_j \in \mathbb{R}^2,\ 1 \le j \le n$ be the rows of design matrix X. Assume $\boldsymbol{z}_j \overset{i.i.d}{\sim} N(0, \rho\mathbf{1}\mathbf{1}^\top + (1-\rho)I_2)$ where $\mathbf{1} = (1,1)^\top$ then:*

(i) ***Mean-Field VI:*** *Let the approximate posterior of interest be $q_{\phi^*}^{MF}$. Then we have:*

$$1-\alpha_{\Pi}^{(i)} \overset{a.e}{\to} 1-2\left\{1-\Phi\left(\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{1-\rho^2}\right)\right\} \text{ as } n\to\infty$$

(ii) ***IAF:*** *Let $q_{\phi^*}^{IAF}$ be the approximate posterior for this case.*

a) ***Case when $K=1$:*** *$\boldsymbol{x_1^\top x_2} = 0 \implies 1-\alpha_{\Pi}^{(i)} = 1-\alpha$. If $\boldsymbol{x_1^\top x_2} \neq 0$, let $q_{\psi^*(\tilde{t})}^{IAF}$ be the sequence of approximate posteriors from Theorem 3.2.1 (ib). When $\boldsymbol{x_1^\top x_2} > 0$, then $\lim_{\tilde{t}\to-\infty} 1-\alpha_{\Pi}^{(i)}(\tilde{t}) = 1-\alpha$ a.e. . If $\boldsymbol{x_1^\top x_2} < 0$, we have the same limits provided $\tilde{t}\to\infty$.*

b) ***Case when $K=2$:*** *$1-\alpha_{\Pi}^{(i)} = 1-\alpha$* [4]

The corollary implies there is no loss in coverage for MF-VI when $\rho = 0$. However, this loss increases monotonically in $|\rho|$ reaching $\approx 0\%$ coverage as $|\rho| \to 1$. There is *almost* no loss in coverage when using $q_{\phi^*}^{\text{IAF}}$ with $K=1$ provided $-\text{sign}(\boldsymbol{x_1^\top x_2})\tilde{t}$ is large enough. There is *absolutely* no loss in coverage when $K=2$ regardless of ρ. Figure 3.2 illustrates that FAVI consistently outperforms MF-VI, achieving $\approx 95\%$ coverage when $\rho > 0,\ \tilde{t} < -2$ and behaves similarly to MF-VI, with near 0% coverage when $\rho > 0,\ \tilde{t} > 2$. The proof of Corollary 3.2.1.1 provides the expression used to plot $1-\alpha_{\Pi}^{(1)}$ in Figure 3.2 (see 3.4.3.3).[5]

[4] Note that while the expressions for limit $\tilde{t}\to\pm\infty$ of coverage $1-\alpha_{\Pi}^{(i)}(\tilde{t})$ when $K=1$ and $1-\alpha_{\Pi}^{(i)}$ when $K=2$ are free of ρ they can be viewed as constant functions of ρ.

[5] The marginal inverse cdf ${F_{\phi^*}^{(2)}}^{-1}$ for θ_2 under IAF ($K=1$) does not admit a closed form. Since $1-\alpha_{\Pi}^{(2)}$ depends on $\tilde{t}$ and the sample correlation $\boldsymbol{x_1^\top x_2}$ via ${F_{\phi^*}^{(2)}}^{-1}$ we need simulations to visualize this relationship. Regardless, for an appropriate $\tilde{t}$, $1-\alpha_{\Pi}^{(2)} \approx 1-\alpha$.

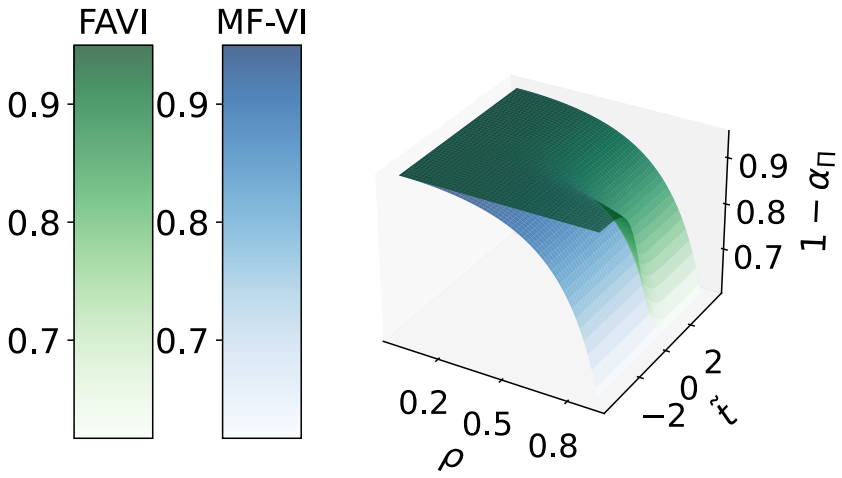


Figure 3.2 Plot of actual coverage $1 - \alpha_{\Pi}^{(1)}$ for θ_1 obtained from the true posterior corresponding to 95% credible intervals from $q_{\phi^*}^{\text{MF}}$ and $q_{\phi^*}^{\text{IAF}}$ when $K = 1$. The x and y axes are ρ and the bias parameter $\tilde{t}$ respectively.

3.3 Summary and Discussion

FAVI has emerged as a powerful tool for Bayesian inference in machine learning. However, some fundamental theoretical gaps in the characterization of this tool still remain. In this chapter, we have used Bayesian linear regression to provide a lens by which the statistical properties of the approximate posterior from FAVI can be analyzed. Our approach goes beyond studying frequentist consistency properties of VI that is the mainstay of current literature, to evaluate the fidelity of the posterior variance estimates from VI. While we have focused on comparisons to MF-VI, we note that a multivariate Gaussian structured variational family Q^{SVI} with covariance $\Sigma \succ 0 \in \mathbb{R}^{2\times 2}$ will contain $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$. In fact, we will see in the next chapter that the family $Q^{\text{IAF}} \supseteq Q^{\text{SVI}}$ and both FAVI with IAF and SVI require $O(p^2)$ parameters to *completely* recover the true posterior.

Substantial future work is necessary to build confidence in FAVI as a tool for statistical inference. In Chapter 4 we will expand our proofs to characterize the behavior of $q_{\phi^*}^{\text{IAF}}$ for $p > 2$ and discuss further potential generalizations of this work.

3.4 Technical Details

In this section we will present the required technical details and proofs for the main results in this chapter. We begin with a glossary of mathematical notation that acts as a supplement to Table

3.1. This will be useful to follow the proof arguments in this section.

3.4.1 Glossary of General Mathematical Notation

Symbol	Description
$KL(q \mid\mid p)$	Kullback-Leibler (KL) divergence measuring "closeness" between 2 distributions. The KL is defined as $KL(q \mid\mid p) = \mathbb{E}_q[\ln \frac{q}{p}]$.
X	Design matrix of predictor variables.
$\boldsymbol{y}$	Observed response vector.
$\boldsymbol{x}_i$	Columns of X.
ρ	Correlation between predictors $\boldsymbol{x}_i, \boldsymbol{x}_j, i \neq j$.
$\boldsymbol{\theta}$	Unknown parameters for statistical inference.
$\boldsymbol{\theta}_0$	Samples from the base distribution q_0 for NFs, $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots \theta_p^0)$.
$\Pi(\boldsymbol{\theta}\|X, \boldsymbol{y})$	Target posterior distribution to sample from.
$m_{\boldsymbol{\theta}},\ \Sigma_{\boldsymbol{\theta}}$	Mean vector and covariance matrix for $\Pi(\boldsymbol{\theta}\|X, \boldsymbol{y})$.
ϵ	Random noise vector.
σ	Standard deviation of ϵ.
$1 - \alpha$	Bayesian credible interval coverage, indicating the probability that θ_i lies in the credible interval $(L_\alpha^{(i)}, U_\alpha^{(i)})$ under the posterior (true posterior or variational posterior).
$\boldsymbol{\theta}_{1:i-1}$ or $\boldsymbol{\theta}_{<i}$	The sub-vector of $\boldsymbol{\theta}$ comprising $(\theta_1, \theta_2, \ldots \theta_{i-1})$ for $2 \leq i \leq p$.
$\boldsymbol{\theta}_{-(i)}$	The sub-vector of $\boldsymbol{\theta}$ excluding θ_i, that is, $(\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_p)$.

Table 3.2 Glossary of General Mathematical Notation used in the manuscript and SI text.

3.4.2 Inverse Auto-Regressive Flows

Symbol	Description
b_i	Translation parameters applied to base distribution samples θ_i^0.
a_i	Scale parameters applied to base distribution samples θ_i^0.
$h(.)$	Any invertible function from $\mathbb{R} \to \mathbb{R}^+$.
c_ϕ^b	Conditioner neural network for IAF whose outputs are b_i.
c_ϕ^a	Conditioner neural network for IAF whose outputs are $h^{-1}(a_i)$.
K	Number of hidden nodes in the conditioner networks c_ϕ.
$\boldsymbol{\theta}_{1:i-1}^0$ or $\boldsymbol{\theta}_{<i}^0$	Sub-vector of first $i-1$ elements of base distribution samples $\boldsymbol{\theta}_0$.
$g(.)$	Activation function (usually refers to the ReLU function $g(x) = \max(x, 0)$).
$\boldsymbol{t}_{\text{in}}^{\{\cdot\}}$	Bias parameter for hidden layer of $c_\phi^{\{\cdot\}}$ ($\cdot$ is a or b).
$\boldsymbol{t}_{\text{out}}^{\{\cdot\}}$	Bias parameter for output layer of $c_\phi^{\{\cdot\}}$ ($\cdot$ is a or b).
$c_{kj}^{\{\cdot\}}$	Entries of weight matrix C for $c_\phi^{\{\cdot\}}$ ($\cdot$ is a or b). We drop subscripts k, j where appropriate when C becomes a vector or scalar due to the mask M^C.
$v_{kj}^{\{\cdot\}}$	Entries of weight matrix V for $c_\phi^{\{\cdot\}}$ ($\cdot$ is a or b). We drop subscripts k, j where appropriate when V becomes a vector or scalar due to the mask M^V.

Table 3.3 Notation pertaining to Inverse Auto-Regressive Flows (IAF) relevant to proofs.

In this section, we present details on Inverse Auto Regressive Flows (IAF) relevant to the proofs presented in the remaining sections. For background information refer to [21].

To generate Q^{IAF} we sample $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots \theta_p^0)^\top \sim N(0, I_p)$ and then apply a transformation $\boldsymbol{\theta} = T(\boldsymbol{\theta}_0)$ as follows:

$$\begin{aligned} &\theta_i = b_i + a_i \theta_i^0, \quad 1 \le i \le p \\ &b_i = c_\phi^b(\theta_{1:i-1}^0) \quad \text{and} \quad a_i = h(c_\phi^a(\theta_{1:i-1}^0)), \quad \text{for } 2 \le i \le p \end{aligned} \tag{3.7}$$

Here $\left\{\{c_\phi^b(\theta_{1:i-1}^0), c_\phi^a(\theta_{1:i-1}^0)\}_{i=2}^p, b_1, a_1\right\}$ comprise the outputs of shallow "conditioner" neural networks $\boldsymbol{c}_\phi^a(.), \boldsymbol{c}_\phi^b(.) : \mathbb{R}^p \to \mathbb{R}^p$; $b_1 \in \mathbb{R}, a_1 \in \mathbb{R}^+$ are free of $\boldsymbol{\theta}_0$ and $h : \mathbb{R} \to \mathbb{R}^+$ is an invertible function that ensures the scaling parameters $a_i > 0$. Keeping with the formulation in the paper first introducing IAF [21], we use conditional masked auto-regressive density estimators (cMADE) for $\boldsymbol{c}_\phi$ [12]. Let $\boldsymbol{\theta}_0 \in \mathbb{R}^p$. We can define the cMADE $\boldsymbol{c}_\phi(.) : \mathbb{R}^p \to \mathbb{R}^p$ as follows:

$$\boldsymbol{c}_\phi(\boldsymbol{\theta}_0) = \boldsymbol{t}_{\text{out}} + (V \odot M^V) g\left(\boldsymbol{t}_{\text{in}} + (C \odot M^C)\boldsymbol{\theta}_0\right) \tag{3.8}$$

Here g is an activation function; $\boldsymbol{t}_{\text{out}} \in \mathbb{R}^p, \boldsymbol{t}_{\text{in}} \in \mathbb{R}^K$ are bias parameters; $V \in \mathbb{R}^{p\times K}, C \in \mathbb{R}^{K\times p}$ are the weight matrices; $M^V \in \mathbb{R}^{p\times K}, M^C \in \mathbb{R}^{K\times p}$ are masking matrices that enforce the auto-regressive structure of c_ϕ and $\odot$ denotes element-wise multiplication.

The masking matrices M^V and M^C have binary $0, 1$ entries and are constructed so as to drop connections between any θ_i and $c_\phi(\boldsymbol{\theta})_j$ for any $i \geq j$. In order to achieve that an integer $m(k)$ is assigned inclusively to each of the hidden nodes in $\boldsymbol{c}_\phi$ such that $1 \leq m(k) < p$. Here, m represents the number of inputs each output of the conditioner depends on. M^V and M^C are then defined as:

$$(M^V)_{ik} = \begin{cases} 1, & \text{if } i > m(k) \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad (M^C)_{ki} = \begin{cases} 1, & \text{if } i \leq m(k) \\ 0, & \text{otherwise} \end{cases} \qquad 1 \leq k \leq K,\ 1 \leq i \leq p.$$

3.4.3 Theoretical Results

Some Preliminary Notation: Section 1.4 has already introduced basic notational conventions used throughout this dissertation. For further simplicity, we use $\tilde{\boldsymbol{x}} = \|\boldsymbol{x}\|_2^2 + 1$ for any vector $\boldsymbol{x} \in \mathbb{R}^m$. Recall that for any scalar $s \in \mathbb{R}$ we use $\text{sign}(s)$ for the sign of s, that is:

$$\text{sign}(s) = \begin{cases} 1, & \text{if } s > 0 \\ -1, & \text{if } s < 0 \\ 0, & s = 0 \end{cases}$$

3.4.3.1 Lemmas

Lemma 3.4.1. *If $Y \sim N(\mu, \sigma^2)$ and $V = \max(Y, 0)$ then V is a rectified gaussian random variable with parameters μ and σ. Its first two moments are given by:*

$$\mathbb{E}[V] = \mu\big(1 - \Phi(-\frac{\mu}{\sigma})\big) + \sigma\phi(\frac{\mu}{\sigma})$$
$$\mathbb{E}[V^2] = \big(\mu^2 + \sigma^2\big)\big(1 - \Phi(-\frac{\mu}{\sigma})\big) + \mu\sigma\phi(\frac{\mu}{\sigma})$$

Proof. See [1] for the proof. □

Lemma 3.4.2. *Let $Y \sim N(0,1)$ and $V = \max(sY + t, 0)$ for some $s, t \in \mathbb{R},\ s \neq 0$. Then:*

$$\mathbb{E}[YV] = s(1 - \Phi(-sign(s)\frac{t}{s}))$$

Proof. Let $\tilde{V} = YV$.

Case (i) $s > 0$: Then $\tilde{V} = \begin{cases} s(Y^2 + Y\frac{t}{s}), & \text{if } Y \geq -\frac{t}{s} \\ 0, & \text{otherwise} \end{cases}$

Therefore,

$$\begin{aligned}
\mathbb{E}[\tilde{V}] &= \mathbb{E}\Big[s\big(Y^2 + Y\frac{t}{s}\big)\mathbb{I}\big\{Y \geq -\frac{t}{s}\big\}\Big] \\
&= s\int_{-\frac{t}{s}}^{\infty} y^2 \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dy + t\int_{-\frac{t}{s}}^{\infty} y\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dy \\
&= s \times \Big(-\frac{1}{s\sqrt{2\pi}} t e^{-\frac{t^2}{2s^2}} + 1 - \Phi(-\frac{t}{s}) + \frac{1}{s\sqrt{2\pi}} t e^{-\frac{t^2}{2s^2}} \Big) \\
&= s\Big(1 - \Phi(-\frac{t}{s})\Big)
\end{aligned}$$

Case (ii) $s < 0$: Then $\tilde{V} = \begin{cases} s(Y^2 + Y\frac{t}{s}), & \text{if } Y \leq -\frac{t}{s} \\ 0, & \text{otherwise} \end{cases}$

$$\begin{aligned}
\mathbb{E}[\tilde{V}] &= \mathbb{E}\Big[s\big(Y^2 + Y\frac{t}{s}\big)\mathbb{I}\big\{Y \leq -\frac{t}{s}\big\}\Big] \\
&= s\int_{-\infty}^{-\frac{t}{s}} y^2 \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dy + t\int_{-\infty}^{-\frac{t}{s}} y\frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}dy
\end{aligned}$$

Substituting $y = -u$ we have,

$$\begin{aligned}
&= s\int_{\infty}^{\frac{t}{s}} -u^2 \frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}du + t\int_{\infty}^{\frac{t}{s}} u\frac{e^{-\frac{u^2}{2}}}{\sqrt{2\pi}}du \\
&= s \times \Big(\frac{1}{s\sqrt{2\pi}} t e^{-\frac{t^2}{2s^2}} + 1 - \Phi(\frac{t}{s}) - \frac{1}{s\sqrt{2\pi}} t e^{-\frac{t^2}{2s^2}} \Big) \\
&= s\Big(1 - \Phi(\frac{t}{s})\Big)
\end{aligned}$$

□

Lemma 3.4.3. *Let $Y \sim N(0,1)$ and $\mu(Y) = \max(sY + t, 0) + b$ for some $s \neq 0,\ t, b \in \mathbb{R}$. Define $V|Y \sim N(\mu(Y), \sigma^2)$ and let $f(v)$ be the marginal distribution of V. Then:*

$$f(v) = \Phi\Big(- sign(s)\frac{t}{s}\Big)\frac{e^{-\frac{(v-b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \Big(1 - \Phi\Big(- sign(s)\Big(\frac{t\sigma^2 + s^2(v-b)}{\sigma s\sqrt{\sigma^2+s^2}}\Big)\Big)\Big)\frac{e^{-\frac{(v-(b+t))^2}{2(\sigma^2+s^2)}}}{\sqrt{2\pi(\sigma^2+s^2)}}$$

Proof. **Case (i) $s > 0$:** First, we note the following.

$$\mu(Y) = \begin{cases} b + sY + t, & \text{if } Y \geq -\frac{t}{s} \\ b, & \text{otherwise} \end{cases}$$

We then have:

$$\begin{aligned}
f(v) &= \int_{\mathbb{R}} f_{V|Y}(v|y)\phi(y)dy = \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(v-\mu(y))^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= \int_{-\infty}^{-\frac{t}{s}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(v-b)^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy + \int_{-\frac{t}{s}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(v-(sy+b+t))^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy \\
&= \Phi\big(- \frac{t}{s}\big)\frac{e^{-\frac{(v-b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \int_{-\frac{t}{s}}^{\infty} \frac{e^{-\frac{(v-(b+t))^2}{2(\sigma^2+s^2)}}}{\sqrt{2\pi}\sigma} \frac{1}{\sqrt{2\pi}} e^{-\frac{\left(y - \frac{s(v-(t+b))}{\sigma^2+s^2}\right)^2(\sigma^2+s^2)}{\sigma^2}} dy \qquad (3.9) \\
&= \Phi\Big(- \frac{t}{s}\Big)\frac{e^{-\frac{(v-b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \Big(1 - \Phi\Big(\frac{-\frac{t}{s} - \frac{s(v-(t+b))}{\sigma^2+s^2}}{\frac{\sigma}{\sqrt{\sigma^2+s^2}}}\Big)\Big)\frac{e^{-\frac{(v-(b+t))^2}{2(\sigma^2+s^2)}}}{\sqrt{2\pi(\sigma^2+s^2)}} \qquad (3.10) \\
&= \Phi\Big(- \frac{t}{s}\Big)\frac{e^{-\frac{(v-b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \Big(1 - \Phi\Big(- \Big(\frac{t\sigma^2 + s^2(v-b)}{\sigma s\sqrt{\sigma^2+s^2}}\Big)\Big)\Big)\frac{e^{-\frac{(v-(b+t))^2}{2(\sigma^2+s^2)}}}{\sqrt{2\pi(\sigma^2+s^2)}}
\end{aligned}$$

(3.9) follows from completing the square and (3.10) follows from properties of ϕ.

Case (i) $s < 0$: For this case, $f(v)$ has the following expression.

$$\begin{aligned}
f(v) &= \int_{-\frac{t}{s}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(v-b)^2}{2\sigma^2}} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy + \int_{-\infty}^{-\frac{t}{s}} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(v-(sy+b+t))^2}{2\sigma^2}} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}} dy \\
&= \Phi\Big(\frac{t}{s}\Big)\frac{e^{-\frac{(v-b)^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} + \Big(1 - \Phi\Big(\Big(\frac{t\sigma^2 + s^2(v-b)}{\sigma s\sqrt{\sigma^2+s^2}}\Big)\Big)\Big)\frac{e^{-\frac{(v-(b+t))^2}{2(\sigma^2+s^2)}}}{\sqrt{2\pi(\sigma^2+s^2)}} \qquad (3.11)
\end{aligned}$$

Where (3.11) is obtained by the same arguments as those of the case $s > 0$. Therefore, the result holds. □

Lemma 3.4.4. *Let $j(.) : \mathbb{R} \to \mathbb{R}$ be defined as $j(\tilde{t}) = g_2(\tilde{t}) - g_1^2(\tilde{t})$ for any $\tilde{t} \in \mathbb{R}$ and fix $s \in \mathbb{R}$. Here, $g_2(\tilde{t}) = (1+\tilde{t}^2)\big(1-\Phi(sign(s)\tilde{t})\big) - sign(s)\tilde{t}\phi(\tilde{t})$ and $g_1(\tilde{t}) = \tilde{t}\big(1-\Phi(sign(s)\tilde{t})\big) - sign(s)\phi(\tilde{t})$. Then:*

(i) $s > 0 \implies \lim_{\tilde{t}\to-\infty} j(\tilde{t}) = 1$

(ii) $s < 0 \implies \lim_{\tilde{t}\to\infty} j(\tilde{t}) = 1$

Proof.

$$
\begin{aligned}
j(\tilde{t}) &= (1+\tilde{t}^2)\big(1-\Phi(\text{sign}(s)\tilde{t})\big) - \text{sign}(s)\tilde{t}\phi(\tilde{t}) - \tilde{t}^2\big(1-\Phi(\text{sign}(s)\tilde{t})\big)^2 \\
&- (\text{sign}(s))^2\phi^2(\tilde{t}) - 2\text{sign}(s)\tilde{t}\big(1-\Phi(\text{sign}(s)\tilde{t})\big)\phi(\tilde{t}) \\
&= \underbrace{\big(1-\Phi(\text{sign}(s)\tilde{t})\big)}_{I} + \underbrace{\tilde{t}^2\big(1-\Phi(\text{sign}(s)\tilde{t})\big)\Phi(\text{sign}(s)\tilde{t})}_{II} \\
&- \underbrace{\text{sign}(s)\tilde{t}\phi(\tilde{t}) - (\text{sign}(s))^2\phi^2(\tilde{t})}_{III} - \underbrace{2\text{sign}(s)\tilde{t}\big(1-\Phi(\text{sign}(s)\tilde{t})\big)\phi(\tilde{t})}_{IV}
\end{aligned}
$$

For the case $s > 0$, it is easy to see that as $\tilde{t} \to -\infty$: $I \to 1$, $III, IV \to 0$ by the properties of ϕ, Φ. For II we can apply L'Hospital's Rule as follows:

$$
\lim_{\tilde{t}\to-\infty} II = \lim_{\tilde{t}\to-\infty} \big(1-\Phi(\tilde{t})\big) \times \lim_{\tilde{t}\to-\infty} \frac{\Phi(\tilde{t})}{t^{-2}} = \lim_{\tilde{t}\to\infty} -\frac{\phi(\tilde{t})t^3}{2} = 0
$$

Therefore when $s > 0$, $\lim_{\tilde{t}\to-\infty} j(\tilde{t}) = 1$. Similar arguments can be used to show that $s < 0 \implies \lim_{\tilde{t}\to\infty} j(\tilde{t}) = 1$. □

Lemma 3.4.5. *Let $u, v \in \mathbb{R}$ s.t $u > v$. Define $j(\tilde{t})$ as in Lemma 3.4.4 and $l(\tilde{t}) = (1 - \Phi(sign(s)\tilde{t}))$, then the following hold:*

(i) $s > 0 \implies \exists\, \tilde{t}_0$ *s.t* $uj(\tilde{t}) - vl^2(\tilde{t}) > 0 \;\forall\, \tilde{t} < \tilde{t}_0$

(ii) $s < 0 \implies \exists\, \tilde{t}_0$ *s.t* $uj(\tilde{t}) - vl^2(\tilde{t}) > 0 \;\forall\, \tilde{t} > \tilde{t}_0$

Proof. To start with let $s > 0$. From Lemma 3.4.4 we know that $\lim_{\tilde{t}\to-\infty} j(\tilde{t}) = 1$ and it is easy to see that $\lim_{\tilde{t}\to-\infty} l^2(\tilde{t}) = 1$. Therefore, $\lim_{\tilde{t}\to-\infty} uj(\tilde{t}) - vl^2(\tilde{t}) = u - v > 0$. The result follows by definition of limits. Similar arguments can be used for the case $s < 0$. □

3.4.3.2 Proof of Main Theorem

For improved readability, we recap some definitions and the theorem statements. We define the Bayesian linear model as follows:

$$\boldsymbol{y} = X\boldsymbol{\theta} + \epsilon, \quad \text{Prior: } \pi(\boldsymbol{\theta}) = \frac{1}{2\pi} e^{-\frac{1}{2}(\theta_1^2+\theta_2^2)} \tag{3.12}$$

Where $\boldsymbol{y} \in \mathbb{R}^n$, $X = [\boldsymbol{x}_1, \boldsymbol{x}_2]$, $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^n$; $\epsilon \sim N(0, \sigma^2 I_n)$ and σ^2 is known.

The true posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ for the model (3.12) above is the $N(m_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$ distribution where $m_{\boldsymbol{\theta}} = \Sigma_{\boldsymbol{\theta}}^{-1} X^\top \boldsymbol{y}$ and $\Sigma_{\boldsymbol{\theta}} = (X^\top X + I_2)^{-1}$. Expanding $m_{\boldsymbol{\theta}}$ as $(m_{\boldsymbol{\theta}1}, m_{\boldsymbol{\theta}2})^\top$ we have:

$$\begin{aligned} m_{\boldsymbol{\theta}1} &= \frac{\tilde{\boldsymbol{x}}_2 \boldsymbol{x}_1^\top \boldsymbol{y} - (\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)\boldsymbol{x}_2^\top \boldsymbol{y}}{\det \Sigma_{\boldsymbol{\theta}}^{-1}} \text{ and } m_{\boldsymbol{\theta}2} = \frac{\tilde{\boldsymbol{x}}_1 \boldsymbol{x}_2^\top \boldsymbol{y} - (\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)\boldsymbol{x}_1^\top \boldsymbol{y}}{\det \Sigma_{\boldsymbol{\theta}}^{-1}} \\ \Sigma_{\boldsymbol{\theta}} &= \frac{1}{D} \begin{bmatrix} \tilde{\boldsymbol{x}}_2 & -\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \\ -\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 & \tilde{\boldsymbol{x}}_1 \end{bmatrix} \text{ Here, } D = \tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)^2 \end{aligned} \tag{3.13}$$

Theorem 3.4.6. *Consider the posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ in (3.13) corresponding to the model (3.12). Let $q_{\phi^*}^{IAF} \in Q^{IAF}$ be the optimal approximate posterior for $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$. Here, Q^{IAF} is the family generated by IAF with K hidden nodes in the shallow network $\boldsymbol{c}_\phi$ from (3.7). That is:*

$$q_{\phi^*}^{IAF} \in argmin_{q_\phi^{IAF} \in Q^{IAF}} KL(q_\phi^{IAF} \,||\, \Pi(.|X, \boldsymbol{y}))$$

Then we have:

(i) ***Case when $K = 1$:***

a) $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \neq 0 \implies KL(q_\phi^{IAF} \,||\, \Pi(.|X, \boldsymbol{y})) > 0$ *for all* $q_\phi^{IAF} \in Q^{IAF}$ *and* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 = 0 \implies KL(q_{\phi^*}^{IAF} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$.

b) $\exists$ *some combination of parameters* $\psi^*(\tilde{t}) = (\phi^*_{-(\tilde{t})}, \tilde{t})$ *s.t* $q_{\psi^*(\tilde{t})}^{IAF}$ *has mean* $m_{\boldsymbol{\theta}}$ *and covariance* $\Sigma_{\psi^*(\tilde{t})}$. *When* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 > 0$ *then* $\lim_{\tilde{t}\to-\infty} \Sigma_{\psi^*(\tilde{t})} = \Sigma_{\boldsymbol{\theta}}$ *element-wise and* $\lim_{\tilde{t}\to-\infty} KL(q_{\psi^*(\tilde{t})}^{IAF} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$. *If* $\boldsymbol{x}_1^\intercal \boldsymbol{x}_2 < 0$, *then we have the same limits provided* $\tilde{t} \to +\infty$.

(ii) ***Case when $K = 2$****: $q^{IAF}_{\phi^*}(\boldsymbol{\theta}) = \Pi(\boldsymbol{\theta}|X, Y)\ \forall\ \boldsymbol{\theta} \in \mathbb{R}^2$ and $KL(q^{IAF}_{\phi^*}\ ||\ \Pi(.|X, \boldsymbol{y})) = 0$.*

Proof. Since σ^2 is known, we can assume without loss of generality that $\sigma^2 = 1$. This is because, if $\sigma^2 \neq 1$ we can use the re-parameterization $\tilde{\boldsymbol{y}} = \sigma^{-1}y$, $\tilde{X} = \sigma^{-1}X$ and $\tilde{\epsilon} = \sigma^{-1}\epsilon$ to write the regression model as $\tilde{\boldsymbol{y}} = \tilde{X}\boldsymbol{\theta} + \tilde{\epsilon}$ for which $\tilde{\epsilon} \sim N(0, I_n)$.

Case when $K = 2$: We begin with this case because it is easier to prove and it will motivate some simplifications we make for the case $K = 1$ later on.

We know that any multivariate normal distribution can be written as the product of the conditional distributions. Therefore:

$$\begin{aligned}
&\Pi(\boldsymbol{\theta}|X, \boldsymbol{y}) = \Pi(\theta_2|\theta_1, X, \boldsymbol{y}) \times \Pi(\theta_1|X, \boldsymbol{y}) \\
&= \frac{1}{\sqrt{2\pi}s_2(\theta_1)} e^{-\frac{1}{2s_2^2(\theta_1)}(\theta_2 - \mu_2(\theta_1))^2} \times \frac{1}{\sqrt{2\pi}s_1} e^{-\frac{1}{2s_1^2}(\theta_1 - \mu_1)^2} \\
&= \frac{1}{\sqrt{2\pi\tilde{\boldsymbol{x}}_2^{-1}}} e^{-\frac{1}{2\tilde{\boldsymbol{x}}_2^{-1}}\left(\theta_2 - (\boldsymbol{x}_2^\top \boldsymbol{y}\tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x_1^\top x_2}\tilde{\boldsymbol{x}}_2^{-1}\theta_1)\right)^2} \times \frac{1}{\sqrt{2\pi\tilde{\boldsymbol{x}}_2 D^{-1}}} e^{-\frac{1}{2\tilde{\boldsymbol{x}}_2 D^{-1}}(\theta_1 - m_{\boldsymbol{\theta}1})^2}
\end{aligned} \tag{3.14}$$

That is, $\theta_1 \sim N(m_{\boldsymbol{\theta}1}, \tilde{\boldsymbol{x}}_2 D^{-1})$ and $\theta_2|\theta_1 \sim N(\mu_2(\theta_1), \tilde{\boldsymbol{x}}_2^{-1})$. Here $\mu_2(\theta_1)$ is given by:

$$\mu_2(\theta_1) = \boldsymbol{x}_2^\top \boldsymbol{y}\tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x_1^\top x_2}\tilde{\boldsymbol{x}}_2^{-1}\theta_1 \tag{3.15}$$

By the properties of KL divergence we have for any two pdfs q and p; $KL(q||p) \geq 0$ and $KL(q||p) = 0 \iff q = p$ a.e. Therefore, if we choose a set of IAF parameters ϕ^* such that $q^{\text{IAF}}_{\phi^*}(\boldsymbol{\theta}) = \Pi(\boldsymbol{\theta}|X, \boldsymbol{y})\ \forall \boldsymbol{\theta} \in \mathbb{R}^2$ then we are done. For convenience, we drop the super-script IAF from q^{IAF}_{ϕ} for the rest of this proof.

The IAF distribution $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim q_\phi$ for $K = 2$ hidden nodes is generated by the following transformations:

$$\begin{aligned}
&\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0) \sim N(0, I_2) \\
&\theta_1 = a_1\theta_1^0 + b_1 \text{ and } \theta_2 = a_2(\theta_1^0) \times \theta_2^0 + b_2(\theta_1^0)
\end{aligned} \tag{3.16}$$

Here $a_1 > 0$, $b_2(\theta_1^0) = \sum_{k=1}^2 g(c_k^b \theta_1^0 + t_{\text{in},k}^b) v_k^b + t_{\text{out}}^b$ and $a_2(\theta_1^0) = h\big(\sum_{k=1}^2 g(c_k^a \theta_1^0 + t_{\text{in},k}^a) v_k^a + t_{\text{out}}^a\big)$ where $g : \mathbb{R} \to \mathbb{R}$ is the ReLU activation, $g(x) = \max\{0, x\}$ and $h : \mathbb{R} \to \mathbb{R}^+$ is an invertible function that ensures $a_2(\theta_1^0) > 0$. In this case:

$$\phi = (a_1, b_1, t_{\text{out}}^a, t_{\text{out}}^b, \{c_k^a, v_k^a, t_{\text{in},k}^a\}_{k=1}^2, \{c_k^b, v_k^b, t_{\text{in},k}^b\}_{k=1}^2)$$

Using the re-parameterization $\theta_1^0 = \frac{\theta_1 - b_1}{a_1}$ we can re-write $b_2(\theta_1^0)$ and $a_2(\theta_1^0)$ as a function of θ_1, that is:

$$b_2(\theta_1^0) = b_2^{'}(\theta_1) = \sum_{k=1}^{2} g\Big(c_k^b \frac{(\theta_1 - b_1)}{a_1} + t_{\text{in},k}^b\Big) v_k^b + t_{\text{out}}^b \tag{3.17}$$

$$a_2(\theta_1^0) = a_2^{'}(\theta_1) = h\bigg(\sum_{k=1}^{2} g\Big(c_k^a \frac{(\theta_1 - b_1)}{a_1} + t_{\text{in},k}^a\Big) v_k^a + t_{\text{out}}^a\bigg) \tag{3.18}$$

We set, $b_1 = m_{\boldsymbol{\theta}1}$, $a_1^2 = \tilde{\boldsymbol{x}}_2 D^{-1}$, $c_k^a = v_k^a = t_{\text{in},k}^a = 0$ for $k = 1, 2$ and $t_{\text{out}}^a = h^{-1}(\sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}})$. Then (3.16), (3.17) and (3.18) imply that $\theta_1 \sim N(m_{\boldsymbol{\theta}1}, \tilde{\boldsymbol{x}}_2 \det \Sigma_{\boldsymbol{\theta}})$ and $\theta_2|\theta_1 \sim N(b_2^{'}(\theta_1), \tilde{\boldsymbol{x}}_2^{-1})$. Therefore we only need to choose $t_{\text{out}}^b, \{c_k^b, v_k^b, t_{\text{in},k}^b\}_{k=1}^2$ s.t $b_2^{'}(\theta_1) = \mu_2(\theta_1)$ in (3.15) and we are done.

Set $t_{\text{in},k}^b = 0$ for $1 \le k \le 2$, $t_{\text{out}}^b = \boldsymbol{x}_2^\top \boldsymbol{y} \tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \tilde{\boldsymbol{x}}_2^{-1} b_1$, $c_1^b = |\boldsymbol{x}_1^\intercal \boldsymbol{x}_2| \tilde{\boldsymbol{x}}_2^{-1} a_1$, $c_2^b = -|\boldsymbol{x}_1^\intercal \boldsymbol{x}_2| \tilde{\boldsymbol{x}}_2^{-1} a_1$, $v_1^b = -\text{sign}(\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)$, $v_2^b = \text{sign}(\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)$. We then have:

$$\begin{aligned}
b_2^{'}(\theta_1) &= \boldsymbol{x}_2^\top \boldsymbol{y} \tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \tilde{\boldsymbol{x}}_2^{-1} b_1 - \text{sign}(\boldsymbol{x}_1^\intercal \boldsymbol{x}_2) |\boldsymbol{x}_1^\intercal \boldsymbol{x}_2| \tilde{\boldsymbol{x}}_2^{-1} a_1 \frac{\theta_1 - b_1}{a_1} \mathbb{I}\{\theta_1 - b_1 \ge 0\} \\
&+ \text{sign}(\boldsymbol{x}_1^\intercal \boldsymbol{x}_2)(-|\boldsymbol{x}_1^\intercal \boldsymbol{x}_2| \tilde{\boldsymbol{x}}_2^{-1} a_1) \frac{\theta_1 - b_1}{a_1} \mathbb{I}\{\theta_1 - b_1 < 0\} \\
&= \boldsymbol{x}_2^\top \boldsymbol{y} \tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \tilde{\boldsymbol{x}}_2^{-1} b_1 - \boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \tilde{\boldsymbol{x}}_2^{-1} (\theta_1 - b_1) \times \big(\mathbb{I}\{\theta_1 - b_1 \ge 0\} + \mathbb{I}\{\theta_1 - b_1 < 0\}\big) \\
&= \boldsymbol{x}_2^\top \boldsymbol{y} \tilde{\boldsymbol{x}}_2^{-1} - \boldsymbol{x}_1^\intercal \boldsymbol{x}_2 \tilde{\boldsymbol{x}}_2^{-1} \theta_1 = \mu_2(\theta_1)
\end{aligned}$$

This completes the proof for $K = 2$.

Case when $K = 1$:

The IAF distribution $\boldsymbol{\theta} = (\theta_1, \theta_2) \sim q_\phi$ for $K = 1$ hidden node is generated by the transformations in (3.16) with $b_2(\theta_1^0) = g(c^b\theta_1^0 + t_{\text{in}}^b)v^b + t_{\text{out}}^b$, $a_2(\theta_1^0) = h\big(g(c^a\theta_1^0 + t_{\text{in}}^a)v^a + t_{\text{out}}^a\big)$ and $g(x) = \max\{0, x\}$.

Let $\boldsymbol{x_1^\intercal x_2} \neq 0$. Note that $KL(q_\phi||\Pi(.|X, \boldsymbol{y})) = 0 \iff q_\phi = \Pi(.|X, \boldsymbol{y})$ a.e. Let if possible $KL(q_\phi||\Pi(.|X, \boldsymbol{y})) = 0$ for some q_ϕ. But if this is true, we know that $q_\phi(\theta_2|\theta_1)$ is gaussian with mean $\mu_2(\theta_1)$. Here $\mu_2(\theta_1)$ is a linear function of θ_1 as in (3.15) and $\mu_2 \neq 0$ when $\boldsymbol{x_1^\intercal x_2} \neq 0$. However, by the IAF transformations (3.16) we have $q_\phi(\theta_2|\theta_1)$ has mean $b_2(\theta_1^0) = b_2(\frac{\theta_1 - b_1}{a_1})$ where b_2 is some non-linear function of θ_1 due to the use of the non-linear activation function g or $b_2 = 0$ when c^b or $v^b = 0$. We arrive at a contradiction.
Therefore, $KL(q_\phi||\Pi(.|X, \boldsymbol{y})) > 0 \; \forall \; q_\phi \in Q^{\text{IAF}}$ when $\boldsymbol{x_1^\intercal x_2} \neq 0$ and $K = 1$.

For the second part of the proof we need to find some combination of parameters $\psi(\tilde{t}) = (\phi^*_{-(\tilde{t})}, \tilde{t})$ such that $KL(q_{\psi(\tilde{t})}||\Pi(.|X, \boldsymbol{y})) \to 0$ as $\tilde{t} \to -\text{sign}(\boldsymbol{x_1^\intercal x_2})\infty$. Here, we have used the following re-parameterization for the input bias parameter t_{in}^b:

$$t_{\text{in}}^b = \tilde{t}c^b \tag{3.19}$$

Motivated by the case $K = 2$, we set:

$$c^a = t_{\text{in}}^a = v^a = 0, \;\; t_{\text{out}}^a = h^{-1}(\sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}}), \;\; v^b = 1 \tag{3.20}$$

This implies $a_2(\theta_1^0)^2 = \frac{1}{\tilde{\boldsymbol{x}}_2}$ and $b_2(\theta_1^0) = g(c^b\theta_1^0 + c^b\tilde{t}) + t_{\text{out}}^b$. Now we are ready to derive the KL divergence as follows:

$$\begin{aligned} &KL(q_\phi||\Pi(|X, \boldsymbol{y})) = \mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[\ln q_\phi(\boldsymbol{\theta}) - \Pi(\boldsymbol{\theta}|X, \boldsymbol{y})\big] \\ &= \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\big[\ln q_\phi(\boldsymbol{\theta}_0)\big] - \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\big[\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})\big] - \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\Big[\ln\Big(\Big|\det\Big(\frac{\partial T}{\partial \boldsymbol{\theta}_0}\Big)\Big|\Big)\Big] \end{aligned} \tag{3.21}$$

Where $q_\phi(\boldsymbol{\theta}_0)$ is the $N(0, I_2)$ distribution and (3.21) follows from the change of variable formula along with properties of expectation [33]. Here, since T refers to the IAF transformation in (3.16),

$\ln\left(\left|\det\left(\frac{\partial T}{\partial \boldsymbol{\theta}_0}\right)\right|\right) = \ln a_1 + \ln a_2 = \ln a_1 + \frac{1}{2}\ln\frac{1}{\tilde{\boldsymbol{x}}_2}$. Continuing we have:

$$
\begin{aligned}
&KL(q_\phi||\Pi(|X,Y)) = \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\Big[-\ln 2\pi - \frac{\theta_1^{0\,2}}{2} - \frac{\theta_1^{0\,2}}{2}\Big] - \frac{1}{2}\ln\det(X^\top X + I_2) + \ln 2\pi \\
&+ \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\Big[\frac{(\boldsymbol{\theta} - (X^\top X + I_2)^{-1}X^\top \boldsymbol{y})^\top (X^\top X + I_2)(\boldsymbol{\theta} - (X^\top X + I_2)^{-1}X^\top \boldsymbol{y})}{2}\Big] \\
&- \ln a_1 - \frac{1}{2}\ln\frac{1}{\tilde{\boldsymbol{x}}_2} \\
&= -1 + \mathbb{E}_{q_0(\boldsymbol{\theta}_0)}\Big[\frac{\tilde{\boldsymbol{x}}_1}{2}\theta_1^2 + \frac{\tilde{\boldsymbol{x}}_2}{2}\theta_2^2 + \boldsymbol{x_1^\top x_2}\theta_1\theta_2 - \boldsymbol{x}_1^\top \boldsymbol{y}\theta_1 - \boldsymbol{x}_2^\top \boldsymbol{y}\theta_2\Big] \\
&- \frac{1}{2}\ln\det(X^\top X + I_2) + \frac{\boldsymbol{y}^\top X(X^\top X + I_2)^{-1}X^\top \boldsymbol{y}}{2} - \ln a_1 - \frac{1}{2}\ln\frac{1}{\tilde{\boldsymbol{x}}_2}
\end{aligned}
$$

Substituting $\theta_1 = a_1\theta_1^0 + b_1, \theta_2 = a_2\theta_2^0 + b_2(\theta_1)$ where $a_2 = \sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}}$ we have:

$$
\begin{aligned}
&KL(q_\phi||\Pi(|X,Y)) = -1 - \frac{1}{2}\ln\det(X^\top X + I_2) + \frac{\boldsymbol{y}^\top X(X^\top X + I_2)^{-1}X^\top \boldsymbol{y}}{2} \\
&- \ln a_1 - \frac{1}{2}\ln\frac{1}{\tilde{\boldsymbol{x}}_2} + \frac{\tilde{\boldsymbol{x}}_1}{2}\big(a_1^2\mathbb{E}_{q_0}[\theta_1^{0\,2}] + b_1^2 + 2a_1b_1\mathbb{E}_{q_0}[\theta_1^0]\big) + \frac{\tilde{\boldsymbol{x}}_2}{2}\big(\frac{1}{\tilde{\boldsymbol{x}}_2}\mathbb{E}_{q_0}[\theta_2^{0\,2}] \\
&+ \mathbb{E}_{q_0}[b_2(\theta_1^0)^2] + 2\sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}} * \mathbb{E}_{q_0}[b_2(\theta_1^0)\theta_2^0]\big) + \boldsymbol{x_1^\top x_2}\mathbb{E}_{q_0}[(a_1\theta_1^0 + b_1)(\sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}}\theta_2^0 + b_2(\theta_1^0))] \\
&- \boldsymbol{x}_1^\top \boldsymbol{y}\mathbb{E}_{q_0}[a_1\theta_1^0 + b_1] - \boldsymbol{x}_2^\top \boldsymbol{y}\mathbb{E}_{q_0}[\sqrt{\frac{1}{\tilde{\boldsymbol{x}}_2}}\theta_2^0 + b_2(\theta_1^0)] \\
&= -1 - \frac{1}{2}\ln\det(X^\top X + I_2) + \frac{\boldsymbol{y}^\top X(X^\top X + I_2)^{-1}X^\top \boldsymbol{y}}{2} - \ln a_1 - \frac{1}{2}\ln\frac{1}{\tilde{\boldsymbol{x}}_2} \\
&+ \frac{\tilde{\boldsymbol{x}}_1}{2}\big(a_1^2 + b_1^2\big) + \frac{\tilde{\boldsymbol{x}}_2}{2}\big(\frac{1}{\tilde{\boldsymbol{x}}_2} + \mathbb{E}_{q_0}[b_2(\theta_1^0)^2]\big) + \boldsymbol{x_1^\top x_2}\big(b_1\mathbb{E}_{q_0}[b_2(\theta_1^0))] + a_1\mathbb{E}_{q_0}[\theta_1^0 b_2(\theta_1^0))]\big) \\
&- \boldsymbol{x}_1^\top \boldsymbol{y}b_1 - \boldsymbol{x}_2^\top \boldsymbol{y}\mathbb{E}_{q_0}[b_2(\theta_1^0)] \qquad (3.22)
\end{aligned}
$$

(3.22) follows from the fact that $\boldsymbol{\theta}_0 \sim N(0, I)$. From Lemmas 3.4.1 and 3.4.2 we know that:

$$\mathbb{E}_{q_0}[b_2(\theta_1^0)] = \kappa_1(c^b, \tilde{t}) + t_{\text{out}}^b \qquad (3.23)$$

$$\mathbb{E}_{q_0}[b_2(\theta_1^0)^2] = \kappa_2(c_b, \tilde{t}) + {t_{\text{out}}^b}^2 + 2t_{\text{out}}^b\kappa_1(c^b, \tilde{t}) \qquad (3.24)$$

$$\mathbb{E}_{q_0}[\theta_1^0 b_2(\theta_1^0))] = \kappa_3(c^b, \tilde{t}) \qquad (3.25)$$

Where:

$$\kappa_1(c^b, \tilde{t}) = c^b \tilde{t}\big(1 - \Phi(-\text{sign}(c^b)\tilde{t})\big) + |c^b|\phi(\tilde{t})$$
$$\kappa_2(c^b, \tilde{t}) = c_b^2(1 + \tilde{t}^2)\big(1 - \Phi(-\text{sign}(c^b)\tilde{t})\big) + c^b|c^b|\tilde{t}\phi(\tilde{t})$$
$$\kappa_3(c^b, \tilde{t}) = c^b(1 - \Phi(-\text{sign}(c^b)\tilde{t}))$$

Although Lemmas 3.4.1 and 3.4.2 apply to the case where $c^b \neq 0$ we see that the above still holds when $c^b = 0$. This is because $c^b = 0 \implies \kappa_1, \kappa_2, \kappa_3 = 0$ and $b_2(\theta_1^0) = t_{\text{out}}^b$. We can now re-write the KL as:

$$h(a_1, b_1, c^b, t_{\text{out}}^b, \tilde{t}) = -\frac{1}{2} - \frac{1}{2}\ln\det(X^\top X + I_2) + \frac{\boldsymbol{y}^\top X(X^\top X + I_2)^{-1}X^\top \boldsymbol{y}}{2}$$
$$- \ln a_1 + \frac{\ln \tilde{\boldsymbol{x}}_2}{2} + \frac{\tilde{\boldsymbol{x}}_1}{2}(a_1^2 + b_1^2) + \frac{\tilde{\boldsymbol{x}}_2}{2}{t_{\text{out}}^b}^2 + \frac{\tilde{\boldsymbol{x}}_2}{2}\kappa_2(c^b, \tilde{t}) + \tilde{\boldsymbol{x}}_2 t_{\text{out}}^b \kappa_1(c^b, \tilde{t})$$
$$+ \boldsymbol{x_1}^\intercal \boldsymbol{x_2} b_1 \kappa_1(c^b, \tilde{t}) + \boldsymbol{x_1}^\intercal \boldsymbol{x_2} b_1 t_{\text{out}}^b + \boldsymbol{x_1}^\intercal \boldsymbol{x_2} a_1 \kappa_3(c^b, \tilde{t}) - \boldsymbol{x}_1^\top \boldsymbol{y} b_1$$
$$- \boldsymbol{x}_2^\top \boldsymbol{y}\kappa_1(c^b, \tilde{t}) - \boldsymbol{x}_2^\top \boldsymbol{y} t_{\text{out}}^b$$

Where $h(.)$ is a real-valued function of multiple variables that represents $KL(q_\phi||\Pi(|X, Y))$. In order to satisfy the result (ib) we choose parameters $(a_1, b_1, c^b, t_{\text{out}}^b)$ as follows:

$$b_1 = m_{\boldsymbol{\theta}1}, \quad a_1^2 = \frac{\tilde{\boldsymbol{x}}_2 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2} \tag{3.26}$$

$$c^b = -a_1 \frac{\boldsymbol{x_1}^\intercal \boldsymbol{x_2}}{\tilde{\boldsymbol{x}}_2}\frac{l(\tilde{t})}{j(\tilde{t})} = \frac{-\boldsymbol{x_1}^\intercal \boldsymbol{x_2} l(\tilde{t})}{\sqrt{(\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2)\tilde{\boldsymbol{x}}_2 j(\tilde{t})}} \tag{3.27}$$

$$t_{\text{out}}^b = \frac{\boldsymbol{x}_2^\top \boldsymbol{y} - \boldsymbol{x_1}^\intercal \boldsymbol{x_2} b_1}{\tilde{\boldsymbol{x}}_2} - \kappa_1(c^b, \tilde{t})$$
$$= m_{\boldsymbol{\theta}2} + \frac{\boldsymbol{x_1}^\intercal \boldsymbol{x_2} l(\tilde{t})}{\sqrt{(\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2)\tilde{\boldsymbol{x}}_2 j(\tilde{t})}} g_1(\tilde{t}) \tag{3.28}$$

Here $j(\tilde{t}),\ l(\tilde{t}),\ g_1(\tilde{t}) : \mathbb{R} \to \mathbb{R}$ are continuous functions of $\tilde{t}$ defined below:

$$j(\tilde{t}) = g_2(\tilde{t}) - g_1^2(\tilde{t}), \quad l(\tilde{t}) = \big(1 - \Phi(\text{sign}(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})\tilde{t})\big) \tag{3.29}$$

$$g_2(\tilde{t}) = (1 + \tilde{t}^2)\big(1 - \Phi(\text{sign}(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})\tilde{t})\big) - \text{sign}(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})\tilde{t}\phi(\tilde{t})$$

$$g_1(\tilde{t}) = \tilde{t}\big(1 - \Phi(\text{sign}(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})\tilde{t})\big) - \text{sign}(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})\phi(\tilde{t}) \tag{3.30}$$

$$\tag{3.31}$$

Note that we need to ensure $a_1 > 0$ to satisfy the IAF parameter constraints.

Case $\boldsymbol{x_1^\top x_2} = 0$: Then $a_1^2 = \frac{1}{\tilde{\boldsymbol{x}}_1} > 0$.

Case $\boldsymbol{x_1^\top x_2} > 0$: By Lemma 3.4.4, $j(\tilde{t}) > 0 \; \forall \; \tilde{t} < \tilde{t}_1$ for some $\tilde{t}_1$. Further, Lemma 3.4.5 and the Cauchy Schwartz inequality imply that $\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2 > 0 \; \forall \; \tilde{t} < \tilde{t}_2$. Thus, our choice of a_1 in (3.26) satisfies: $a_1 > 0 \; \forall \; \tilde{t} < \tilde{t}_0 = \min(\tilde{t}_1, \tilde{t}_2)$.

Case $\boldsymbol{x_1^\top x_2} < 0$: Similar arguments for this case leads us to $a_1 > 0 \; \forall \; \tilde{t} > \tilde{t}_0$ for some $\tilde{t}_0 \in \mathbb{R}$.

We restrict our choice of a_1 to values $\tilde{t} < \tilde{t}_0$ or $\tilde{t} > \tilde{t}_0$ depending on the sign of $\boldsymbol{x_1^\top x_2}$.

Substituting the parameters from (3.26)-(3.28) we can reduce h to:

$$h(\tilde{t}) = -\frac{1}{2}\ln j(\tilde{t}) + \frac{1}{2}\ln \frac{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 - (\boldsymbol{x_1^\top x_2})^2}$$

Clearly, $\boldsymbol{x_1^\top x_2} = 0 \implies \forall \; \tilde{t} \in \mathbb{R}, \; h(\tilde{t}) = \min_{q_{\phi^*}} KL(q_{\phi^*} || \Pi(.|X, \boldsymbol{y})) = 0$. In fact, since $c^b = 0$ for this case, $\theta_1 \perp\!\!\!\perp \theta_2$ and the resulting distribution q_ϕ belongs to the mean-field variational family. It stands to reason that when there is no dependency between the predictors $\boldsymbol{x}_i$ a mean-field family is sufficient to characterize our posterior distribution.

For the remaining cases, observe that by Lemma 3.4.4 and the properties of $\Phi(.)$, $\lim_{\tilde{t}\to-\infty} h(\tilde{t}) = 0$ when $\boldsymbol{x_1^\top x_2} > 0$ and $\lim_{\tilde{t}\to\infty} h(\tilde{t}) = 0$ when $\boldsymbol{x_1^\top x_2} < 0$.

Now it only remains to derive the mean and covariance for $q_{\psi^*(\tilde{t})}$ where $\psi^*(\tilde{t})$ represents the parameters selected in (3.20), (3.26)-(3.28).

$$\text{Trivially, } \mathbb{E}[\theta_1] = m_{\boldsymbol{\theta}1} \text{ and } \text{Var}[\theta_1] = \frac{\tilde{\boldsymbol{x}}_2 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2}$$

$$\mathbb{E}[\theta_2] = \mathbb{E}\big[\mathbb{E}[\theta_2|\theta_1]\big] = \mathbb{E}[b_2(\theta_1^0)] = t_{\text{out}}^b + \kappa_1(c^b, \tilde{t}) = m_{\boldsymbol{\theta}2}$$

$$\text{Var}[\theta_2] = \mathbb{E}[\theta_2^2] - \mathbb{E}[\theta_2]^2 = a_2^2 + \mathbb{E}[b_2(\theta_1^0)^2] - (t_{\text{out}}^b + \kappa_1(c^b, \tilde{t}))^2$$
$$= a_2^2 + {t_{\text{out}}^b}^2 + 2t_{\text{out}}^b \kappa_1(c^b, \tilde{t}) - {t_{\text{out}}^b}^2 - \kappa_1(c^b, \tilde{t})^2 - 2t_{\text{out}}^b \kappa_1(c^b, \tilde{t})$$
$$= \frac{1}{\tilde{\boldsymbol{x}}_2} + \frac{(\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2}{\tilde{\boldsymbol{x}}_2(\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2)} = \frac{\tilde{\boldsymbol{x}}_1 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2}$$

$$\text{Cov}[\theta_1, \theta_2] = \mathbb{E}[\theta_1\theta_2] - \mathbb{E}[\theta_1]\mathbb{E}[\theta_2] = a_1\mathbb{E}[\theta_1^0 b_2(\theta_1^0)] + b_1\mathbb{E}[b_2(\theta_1^0)] - b_1\mathbb{E}[b_2(\theta_1^0)]$$
$$= a_1 c^b l(\tilde{t}) = \frac{-\boldsymbol{x_1}^\intercal \boldsymbol{x_2} l(\tilde{t})^2}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2}$$

The equations above follow from (3.23),(3.24), the IAF transformations in (3.16) and the subsitution of the optimal parameters selected in (3.20), (3.26)-(3.28).

Therefore, we have $q_{\psi^*(\tilde{t})}$ has mean $m_{\boldsymbol{\theta}}$, the same as the true posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ and the covariance matrix is given by:

$$\Sigma_{\phi^*_{-(\tilde{t})}, \tilde{t}} = \begin{bmatrix} \frac{\tilde{\boldsymbol{x}}_2 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2} & \frac{-\boldsymbol{x_1}^\intercal \boldsymbol{x_2} l(\tilde{t})^2}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2} \\ \frac{-\boldsymbol{x_1}^\intercal \boldsymbol{x_2} l(\tilde{t})^2}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2} & \frac{\tilde{\boldsymbol{x}}_1 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1}^\intercal \boldsymbol{x_2})^2 l(\tilde{t})^2} \end{bmatrix} \tag{3.32}$$

Clearly Lemma 3.4.4 and definition of $l(\tilde{t})$ imply that for $1 \leq i, j \leq 2$, $\lim_{\tilde{t}\to-\infty} \left(\Sigma_{\psi^*(\tilde{t})}\right)_{ij} = \left(\Sigma_{\boldsymbol{\theta}}\right)_{ij}$ when $\boldsymbol{x_1}^\intercal \boldsymbol{x_2} > 0$ and $\lim_{\tilde{t}\to\infty} \left(\Sigma_{\psi^*(\tilde{t})}\right)_{ij} = \left(\Sigma_{\boldsymbol{\theta}}\right)_{ij}$ when $\boldsymbol{x_1}^\intercal \boldsymbol{x_2} < 0$. $\Sigma_{\boldsymbol{\theta}}$ is the true posterior covariance matrix as in (3.13). Therefore, the result holds.

Some Remarks:

Our choice of IAF parameters in (3.26)-(3.28) is motivated by the solution to the first order gradient equations $\frac{\partial h}{\partial a_1} = 0, \frac{\partial h}{\partial b_1} = 0, \ldots \frac{\partial h}{\partial t_{\text{out}}^b} = 0$. The idea is that in order to minimize the function $h(.)$, the first order gradient conditions must be satisfied. However, we note that the results ia and ib of Theorem 3.4.6 imply that the global minimum $KL(q_{\phi^*}||\Pi(.|X, \boldsymbol{y}))$ does not exist when $\boldsymbol{x_1}^\intercal \boldsymbol{x_2} \neq 0$. In general, the stochastic gradient descent algorithm used for minimizing the KL will yield reasonably well performing local minima.

□

3.4.3.3 Proof of Corollary

Some Preliminaries: $F^{(i)}_{\phi^*}$ denotes the cdf for $q_{i,\phi^*} = \int_{\theta_{-(i)}} q_{\phi^*}(\boldsymbol{\theta}) d\theta_{-(i)}$ then the $1-\alpha$ equal-tailed credible interval for θ_i is given by:

$$(F^{(i)^{-1}}_{\phi^*}(\frac{\alpha}{2}), F^{(i)^{-1}}_{\phi^*}(1-\frac{\alpha}{2})) \tag{3.33}$$

Let $1-\alpha^{(i)}_{\Pi}$ be the **actual** coverage for this credible interval under the true marginal posterior $\Pi_i(\theta_i|X,\boldsymbol{y}) = \int_{\theta_{-(i)}} \Pi(\boldsymbol{\theta}|X,\boldsymbol{y}) d\theta_{-(i)}$. Then $1-\alpha^{(i)}_{\Pi}$ satisfies:

$$1-\alpha^{(i)}_{\Pi} = \Phi\Big(\frac{F^{(i)^{-1}}_{\phi^*}(1-\frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}^{\frac{1}{2}}}\Big) - \Phi\Big(\frac{F^{(i)^{-1}}_{\phi^*}(\frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}^{\frac{1}{2}}}\Big) \tag{3.34}$$

The difference $(1-\alpha) - (1-\alpha^{(i)}_{\Pi})$ represents the loss in uncertainty quantification when replacing the true posterior with an approximation.

Corollary 3.4.6.1. *Let $1-\alpha$ be the coverage for the equal-tailed credible intervals for $\theta_i,\ i=1,2$ specified in (3.33), corresponding to the approximate posterior q_{ϕ^*}. $1-\alpha^{(i)}_{\Pi}$ defined in (3.34) is the actual coverage for this interval provided by the true posterior. Let $\boldsymbol{z}_j \in \mathbb{R}^2,\ 1 \le j \le n$ be the rows of design matrix X. Assume $\boldsymbol{z}_j \overset{i.i.d}{\sim} N(0, \rho \mathbf{1}\mathbf{1}^\top + (1-\rho) I_2)$ where $\mathbf{1} = (1,1)^\top$, then the following hold:*

(i) ***Mean-Field VI:*** *Let the approximate posterior of interest be $q^{MF}_{\phi^*}$. Then we have:*

$$1-\alpha^{(i)}_{\Pi} \overset{a.e}{\to} 1 - 2\Big\{1 - \Phi\Big(\Phi^{-1}(1-\frac{\alpha}{2})\sqrt{1-\rho^2}\Big)\Big\} \text{ as } n \to \infty$$

(ii) ***IAF:*** *The approximate posterior for this case is $q^{IAF}_{\phi^*}$.*

a) ***Case when $K = 1$:*** *$\boldsymbol{x}_1^\top \boldsymbol{x}_2 = 0 \implies 1-\alpha^{(i)}_{\Pi} = 1-\alpha$. If $\boldsymbol{x}_1^\top \boldsymbol{x}_2 \neq 0$, let $q^{IAF}_{\psi^*(\tilde{t})}$ be the sequence of approximate posteriors from Theorem 3.4.6 (ib). Then, $\lim_{\tilde{t} \to -sign(\boldsymbol{x}_1^\top \boldsymbol{x}_2)\infty} 1-\alpha^{(i)}_{\Pi}(\tilde{t}) = 1-\alpha$ a.e.*

b) ***Case when $K = 2$:*** *$1-\alpha^{(i)}_{\Pi} = 1-\alpha$*

Proof. **Case i MF-VI:** Recall that for the case of MF-VI, $\theta_1 \perp\!\!\!\perp \theta_2$ for any $q^{\text{MF}}_{\phi^*} \in Q^{\text{MF}}$. We define:

$$Q^{\text{MF}} = \left\{ \frac{1}{2\pi} e^{-\sum_{i=1}^{2} \frac{(\theta_i - \mu_i)^2}{2s_i^2}} \,\middle|\, \mu_i \in \mathbb{R}, s_i > 0 \in \mathbb{R},\ i = 1, 2 \right\}$$

The KL divergence between any member of the mean-field family and the true posterior is then given by:

$$\begin{aligned} KL(q^{\text{MF}}_{\phi} || \Pi(.|X, \boldsymbol{y})) = & \frac{1}{2} \Big\{ \sum_{i=1}^{2} \tilde{\boldsymbol{x}}_i s_i^2 - 2 - \ln(\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x_1}^\top \boldsymbol{x_2})^2) + \sum_{i=1}^{2} - \ln s_i^2 \\ & + (m_{\boldsymbol{\theta}} - \mu)^\top (X^\top X + I)(m_{\boldsymbol{\theta}} - \mu) \Big\} \end{aligned}$$

$m_{\boldsymbol{\theta}}$ and $(X^\top X + I_2)^{-1}$ are the mean and covariance of the true posterior respectively. See (3.13).

It is easy to see from the first and second order gradient conditions that the KL is minimized when $\mu = m_{\boldsymbol{\theta}}$ and $s_i^2 = \frac{1}{\tilde{\boldsymbol{x}}_i}$ for $i = 1, 2$. Thus, the optimal $q^{\text{MF}}_{\phi^*}$ has the distribution $\theta_1 \sim N(m_{\boldsymbol{\theta}1}, \frac{1}{\tilde{\boldsymbol{x}}_1}) \perp\!\!\!\perp \theta_2 \sim N(m_{\boldsymbol{\theta}2}, \frac{1}{\tilde{\boldsymbol{x}}_2})$.

Now, a $1 - \alpha$ equal tailed credible interval obtained from $q^{\text{MF}}_{\phi^*}$ for $\theta_i,\ i = 1, 2$ is given by:

$$m_{\boldsymbol{\theta}i} \pm \Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{1}{\tilde{\boldsymbol{x}}_i}}$$

The actual coverage $1 - \alpha^{(i)}_{\Pi}$ provided by the true posterior corresponding to the credible interval above can be obtained from (3.34). In this case, both the approximate posterior and true posterior are gaussian with the same mean $m_{\boldsymbol{\theta}}$. Consequently, we can solve for $\alpha^{(i)}_{\Pi}$ by equating the length of the credible intervals obtained from $q^{\text{MF}}_{\phi^*}$ and $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ as follows:

$$\begin{aligned} & \Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{1}{\tilde{\boldsymbol{x}}_i}} = \Phi^{-1}(1 - \frac{\alpha^{(i)}_{\Pi}}{2}) \sqrt{\frac{\tilde{\boldsymbol{x}}_j}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x_1}^\top \boldsymbol{x_2})^2}}, \text{ for } j \neq i \\ \implies & \alpha^{(i)}_{\Pi} = 2 \left\{ 1 - \Phi\left(\Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x_1}^\top \boldsymbol{x_2})^2}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2}} \right) \right\} \end{aligned}$$

We have assumed that the rows of the design matrix are distributed as $\boldsymbol{z}_j = (z_{j1}, z_{j2})^\top \overset{i.i.d}{\sim} N(0, \rho \mathbf{1}\mathbf{1}^\top + (1 - \rho) I_2)$ for $1 \leq j \leq n$. Combining this with the Strong Law of Large Numbers

(S.L.L.N) we see that:

$$\frac{\tilde{\boldsymbol{x}}_i}{n} = \frac{\|\boldsymbol{x}_i\|_2^2 + 1}{n} = \sum_{j=1}^{n} \frac{z_{ji}^2}{n} + \frac{1}{n} \xrightarrow{\text{a.e}} 1 \quad \text{for } n \to \infty,\ i = 1, 2 \tag{3.35}$$

$$\frac{\boldsymbol{x}_1^\top \boldsymbol{x}_2}{n} = \sum_{j=1}^{n} \frac{z_{j1} z_{j2}}{n} \xrightarrow{\text{a.e}} \rho \quad \text{for } n \to \infty \tag{3.36}$$

$$\implies \sqrt{\frac{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x}_1^\top \boldsymbol{x}_2)^2}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2}} \xrightarrow{\text{a.e}} \sqrt{1 - \rho^2} \quad \text{as } n \to \infty$$

Therefore, by continuity of Φ and properties of limits $n \to \infty \implies$:

$$1 - \alpha_{\Pi}^{(i)} \xrightarrow{\text{a.s}} 1 - 2\left\{1 - \Phi\left(\Phi^{-1}(1 - \frac{\alpha}{2})\sqrt{1 - \rho^2}\right)\right\}$$

Case ii IAF $K = 1$: The result for the case when $\boldsymbol{x}_1^\top \boldsymbol{x}_2 = 0$ follows from Theorem 3.4.6 (ia), that is, $\boldsymbol{x}_1^\top \boldsymbol{x}_2 = 0 \implies KL(q_{\phi^*}^{\text{IAF}} || \Pi(.|X, \boldsymbol{y})) = 0$.

Case when $\boldsymbol{x}_1^\top \boldsymbol{x}_2 > 0$: Let $q_{\psi^*(\tilde{t}_m)}^{\text{IAF}}$ be a sequence of approximate posterior distributions generated by IAF flows with $K = 1$ hidden node that satisfies $KL(q_{\psi^*(\tilde{t}_m)}^{\text{IAF}} || \Pi(.|X, \boldsymbol{y})) \to 0$ as $\tilde{t}_m \to -\infty$. We know that such a sequence exists due to Theorem 3.4.6.

Let $F_{\tilde{t}_m}$ be the sequence of cdfs corresponding to $q_{\psi^*(\tilde{t}_m)}^{\text{IAF}}$. We know that, convergence in KL divergence $\implies$ weak convergence as a consequence of Pinsker's inequality and definition of total variation distance. As a result of this:

$$\tilde{t}_m \to -\infty \implies F_{\tilde{t}_m}^{(i)} \xrightarrow{\text{weakly}} F_{\Pi}^{(i)} \text{ for } i = 1, 2 \tag{3.37}$$

Here $F_{\tilde{t}_m}^{(i)}$ is the marginal cdf for $q_{i,\psi^*(\tilde{t}_m)}^{\text{IAF}} = \int_{\theta_{-(i)}} q_{\psi^*(\tilde{t}_m)}^{\text{IAF}}(\boldsymbol{\theta}) d\theta_{-(i)}$ and $F_{\Pi}^{(i)}$ is the marginal cdf for $\Pi_i(\theta_i|X, \boldsymbol{y})$.

From (3.34) we know that the actual coverage $1 - \alpha_{\Pi}^{(i)}$ is then given by:

$$1 - \alpha_{\Pi}^{(i)} = \Phi\left(\frac{F_{\tilde{t}_m}^{(i)^{-1}}(1 - \frac{\alpha}{2}) - m_{\boldsymbol{\theta}i}}{\left(\Sigma_{\boldsymbol{\theta}}\right)_{ii}^{\frac{1}{2}}}\right) - \Phi\left(\frac{F_{\tilde{t}_m}^{(i)^{-1}}(\frac{\alpha}{2}) - m_{\boldsymbol{\theta}i}}{\left(\Sigma_{\boldsymbol{\theta}}\right)_{ii}^{\frac{1}{2}}}\right) \tag{3.38}$$

By continuity of Φ and properties of limits we have:

$$\begin{aligned}
\lim_{\tilde{t}_m \to -\infty} 1 - \alpha_{\Pi}^{(i)} &= \Phi\Big(\lim_{\tilde{t}_m \to -\infty} \frac{F_{\tilde{t}_m}^{(i)^{-1}}(1 - \frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}^{\frac{1}{2}}} \Big) - \Phi\Big(\lim_{\tilde{t}_m \to -\infty} \frac{F_{\tilde{t}_m}^{(i)^{-1}}(\frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}^{\frac{1}{2}}} \Big) \\
&= \Phi\Big(\frac{m_{\boldsymbol{\theta} i} + \big(\Sigma_{\boldsymbol{\theta}}\big)_{ii} \Phi^{-1}(1 - \frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}} \Big) - \Phi\Big(\frac{m_{\boldsymbol{\theta} i} - \big(\Sigma_{\boldsymbol{\theta}}\big)_{ii} \Phi^{-1}(1 - \frac{\alpha}{2}) - m_{\boldsymbol{\theta} i}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{ii}} \Big) \\
&= 1 - \frac{\alpha}{2} - \big(1 - \Phi\big(\Phi^{-1}(1 - \frac{\alpha}{2})\big)\big) = 1 - \alpha
\end{aligned}$$

Case when $\boldsymbol{x_1^\top x_2 < 0}$: It is easy to see that when $\boldsymbol{x_1^\top x_2} < 0$, taking $\tilde{t}_m \to \infty$ leads to the same result.

Case iii IAF $\boldsymbol{K = 2}$: Follows trivially from Theorem 3.4.6 (ii).

Some Remarks:

We have shown for the case $K = 1$ that $\lim_{\tilde{t}_m \to -\text{sign}(\boldsymbol{x_1^\top x_2})\infty} 1 - \alpha_{\Pi}^{(i)} = 1 - \alpha$. However, this result holds only in a limiting sense and does not explicitly characterize the coverage $1 - \alpha_{\Pi}^{(i)}$ as a function of $\tilde{t}$ and the correlation ρ between predictors $\boldsymbol{x}_i$. We provide some comments on this below.

From Theorem 3.4.6 we know that $\theta_1 \sim N(m_{\boldsymbol{\theta} 1}, \frac{\tilde{\boldsymbol{x}}_2 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2})$ where $j(.)$, $l(.)$ are functions defined in (3.29). Therefore the $1 - \alpha$ credible interval for θ_1 is given by $m_{\boldsymbol{\theta} 1} \pm \Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{\tilde{\boldsymbol{x}}_2 j(\tilde{t})}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2}}$. Following the steps in the MF-VI case we then have:

$$\alpha_{\Pi}^{(1)} = 2\Bigg\{1 - \Phi\Bigg(\Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{\big(\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 - (\boldsymbol{x_1^\top x_2})^2\big) j(\tilde{t})}{\tilde{\boldsymbol{x}}_1 \tilde{\boldsymbol{x}}_2 j(\tilde{t}) - (\boldsymbol{x_1^\top x_2})^2 l(\tilde{t})^2}}\Bigg)\Bigg\}$$

By (3.35) and (3.36) $\alpha_{\Pi}^{(1)} \overset{\text{a.e}}{\to} 2\Big\{1 - \Phi\Big(\Phi^{-1}(1 - \frac{\alpha}{2}) \sqrt{\frac{(1-\rho^2) j(\tilde{t})}{j(\tilde{t}) - l(\tilde{t})^2 \rho^2}}\Big)\Big\}$ as $n \to \infty$. Therefore, we can explicitly calculate $\alpha_{\Pi}^{(1)}$ as a function of $\tilde{t}$ and ρ. Note that the expression above is valid only for $\tilde{t} < \tilde{t}_0 \in \mathbb{R}$ if $\boldsymbol{x_1^\top x_2} > 0$ and $\tilde{t} > \tilde{t}_0 \in \mathbb{R}$ if $\boldsymbol{x_1^\top x_2} < 0$ where $\tilde{t}_0$ is selected to ensure $q_{\psi^*(\tilde{t})}^{\text{IAF}}$ is a valid distribution. See the proof of Theorem 3.4.6 for details.

For calculating $\alpha_{\Pi}^{(2)}$ the same arguments will not work since the marginal $q^{\text{IAF}}_{\psi^*(\tilde{t})}(\theta_2)$ is not gaussian. We instead have:

$$1-\alpha_{\Pi}^{(2)} = \Phi\Big(\frac{F_{\tilde{t}}^{(2)^{-1}}(1-\frac{\alpha}{2}) - m_{\boldsymbol{\theta}2}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{22}{}^{\frac{1}{2}}}\Big) - \Phi\Big(\frac{F_{\tilde{t}}^{(2)^{-1}}(\frac{\alpha}{2}) - m_{\boldsymbol{\theta}2}}{\big(\Sigma_{\boldsymbol{\theta}}\big)_{22}{}^{\frac{1}{2}}}\Big)$$

From Lemma 3.4.3 we know that $F_{\tilde{t}}^{(2)}$ is the cdf for a random variable with probability distribution:

$$f(v) = \Phi\Big(\text{sign}(\boldsymbol{x_1}^{\intercal}\boldsymbol{x_2})\tilde{t}\Big)\sqrt{\frac{\tilde{\boldsymbol{x}}_2}{2\pi}}e^{-\frac{(v-b)^2\tilde{\boldsymbol{x}}_2}{2}} + \Bigg\{\Big(1-\Phi\Big(\text{sign}(\boldsymbol{x_1}^{\intercal}\boldsymbol{x_2})\Big(\frac{\tilde{t}c^b\tilde{\boldsymbol{x}}_2 + c^{b^2}(v-b)}{\sqrt{\tilde{\boldsymbol{x}}_2}c^b\sqrt{\tilde{\boldsymbol{x}}_2 + c^{b^2}}}\Big)\Big)\Big)$$
$$\times \frac{e^{-\frac{(v-(b+\tilde{t}c^b))^2}{2(\tilde{\boldsymbol{x}}_2+c^{b^2})}}}{\sqrt{2\pi(\tilde{\boldsymbol{x}}_2 + c^{b^2})}}\Bigg\}$$

Where $b = m_{\theta_2} - c^b g_1(\tilde{t})$, c^b and $g_1(.)$ are defined in (3.27) and (3.30) respectively. The cdf $F_{\tilde{t}}^{(2)}$ does not have an analytic expression and we have to resort to simulations to approximate the behaviour of $1-\alpha_{\Pi}^{(2)}$ at various values of $\tilde{t}$ and ρ. □

3.5 Extensions to higher dimensions

In the previous sections, we characterized the behaviour of the optimal variational posterior $q^{\text{IAF}}_{\phi^*}$ resulting from using FAVI with IAF, for the Bayesian linear regression model with gaussian priors and exactly two predictors ($p = 2$). A major goal of ours was to provide a concrete, analytical formulation for the loss in credible interval coverage from using the FAVI/MF-VI approximation in place of the true posterior; across complexity levels of the IAF transformation ($K = 1, 2$).[6] To this end, we made the simplifying assumption of a covariate dimension $p = 2$. This assumption (among others), allowed us to get a closed form KL divergence and model the covariance matrix as a function of a single parameter ρ (the correlation between the two covariates). As a result, we were able to achieve the desired goal.

We will now consider extensions to the case $p > 2$. Given the increased complexity of this set-up, we will have to relax our requirements for a very explicit characterization of the optimal KL

[6]We do not consider $K > 2$ since our results showed that $K = 2$ is sufficient to completely recover the posterior and using $K > 2$ would result in an over-parameterized variational family.

divergence and loss in credible interval coverage, and derive upper bounds instead. Generalizations to non-Gaussian posteriors, more complex NF families and non-linear activations are reserved for future work.

Recollect that we use the number of hidden nodes in the conditioner network c_ϕ as the measure of complexity of the IAF transformation. We will focus on answering two questions:

(i) What is the IAF complexity required for FAVI to completely recover the true posterior?

(ii) If K^* denotes the IAF complexity required to completely recover the true posterior, what is the loss in statistical accuracy if we use complexity $K < K^*$?

How to measure statistical accuracy in this context is an open ended question. Since the variational posterior is obtained by minimizing the KL divergence (See 3.1), we derive an upper bound for $\min_{q_\phi^{\text{IAF}} \in Q^{\text{IAF}}} KL(q_\phi \,||\, \Pi(\boldsymbol{\theta}|D))$ as a starting point. For a recap of frequently used mathematical notation see Tables 3.1 and 3.2.

3.5.1 Main Result

We first present some necessary preludes and then state the main result. The Bayesian linear model we consider is defined as:

$$\boldsymbol{y} = X\boldsymbol{\theta} + \epsilon, \quad \text{Prior: } \pi(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}^p} e^{-\frac{1}{2}\|\boldsymbol{\theta}\|_2^2} \tag{3.39}$$

Where $\boldsymbol{y} \in \mathbb{R}^n$, $X = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p]$, $\boldsymbol{x}_i \in \mathbb{R}^n$ for $1 \leq i \leq p$; $\epsilon \sim N(0, \sigma^2 I_n)$ and σ^2 is known. The true posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ for the model (3.39) above is the $N(m_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$ distribution where:

$$m_{\boldsymbol{\theta}} = \Sigma_{\boldsymbol{\theta}}^{-1} X^\top \boldsymbol{y} \;\text{ and }\; \Sigma_{\boldsymbol{\theta}} = (X^\top X + I_p)^{-1} \tag{3.40}$$

Let us denote the k^{th} largest eigen value of $\Sigma_{\boldsymbol{\theta}}$ by γ_k for $1 \leq k \leq p$. We know that $\Sigma_{\boldsymbol{\theta}}$ is a positive definite matrix, since its inverse is the sum of a positive semi-definite matrix $X^\top X$ and the identity matrix I_p. Therefore its minimum eigen value $\lambda_{\min}(\Sigma_{\boldsymbol{\theta}}) = \gamma_p > 0$. We are now ready to state the main result.

Theorem 3.5.1. *Consider the posterior* $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ *defined in (3.40). Let* $q_{\phi^*}^{IAF} \in Q^{IAF}$ *be the optimal approximate posterior for* $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$. *Here,* Q^{IAF} *is the family generated by IAF with K hidden nodes in the shallow conditioner network* c_ϕ *defined in (3.4). That is:*

$$q_{\phi^*}^{IAF} \in argmin_{q_\phi^{IAF} \in Q^{IAF}} KL(q_\phi^{IAF} \,||\, \Pi(.|X, \boldsymbol{y}))$$

Let $\gamma_1 \geq \gamma_2 \cdots \geq \gamma_p$ *be the eigen values of the covariance matrix* $\Sigma_{\boldsymbol{\theta}}$. *We assume* $K = 2K^{'}$ *where* $1 \leq K^{'} < p - 1$. *Then we have:*

$$\min_{q_\phi^{IAF} \in Q^{IAF}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \sum_{k=K'+2}^{p} (\gamma_k - \gamma_p)^2 \tag{3.41}$$

First observe that when $K^{'} = (p-2)$ or equivalently $K = 2(p-2)$, we have:

$$\min_{q_\phi^{\text{IAF}} \in Q^{\text{IAF}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2}(\gamma_p - \gamma_p)^2 = 0$$

Therefore, from the theorem we know that in order for FAVI with IAF to exactly recover the true posterior, i.e. $KL(q_{\phi^*}^{\text{IAF}} \,||\, \Pi(.|X, \boldsymbol{y})) = 0$, we require $K = 2(p-2)$ hidden nodes in the conditioner network c_ϕ. It naturally follows that the corresponding loss in credible interval coverage from the IAF approximation would be 0%.

Since we are working with a Gaussian target posterior, completely characterizing $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ boils down to recovering the mean $m_{\boldsymbol{\theta}}$ and covariance matrix $\Sigma_{\boldsymbol{\theta}}$. Theorem 3.5.1 tell us that for the remaining cases $K < 2(p-2)$, the approximate posterior is able to explain the covariance matrix $\Sigma_{\boldsymbol{\theta}}$ upto its first $K^{'} + 1 = K/2 + 1$ eigen values. The remaining unexplained component is shaped by the magnitude of the $p - K/2 - 1$ smallest eigen values of $\Sigma_{\boldsymbol{\theta}}$ after subtracting the minimum eigen value from them.

3.5.2 Technical Details

Some Preliminaries: For any vector $\boldsymbol{x} = (x_1, x_2, \ldots x_d)^\top \in \mathbb{R}^d$ we denote $\boldsymbol{x}_{i:j} = (x_i, x_{i+1}, \ldots x_j)$, $i < j$ to be the sub-vector of $\boldsymbol{x}$ running from the i^{th} to the j^{th} element. Similarly, let x_{ij} be the ij^{th}

element of matrix $X \in \mathbb{R}^{m\times d}$ and $X_{i_1:i_2,j_1:j_2} \in \mathbb{R}^{(i_2-i_1+1)\times(j_2-j_1+1)}$ be the sub-matrix of X given by:

$$\begin{bmatrix} x_{i_1,j_1} & x_{i_1,j_1+1} & \dots & x_{i_1,j_2} \\ x_{i_1+1,j_1} & x_{i_1+1,j_1+1} & \dots & x_{i_1+1,j_2} \\ \vdots & \vdots & \vdots & \\ x_{i_2,j_1} & x_{i_2,j_1+1} & \dots & x_{i_2,j_2} \end{bmatrix}$$

We denote the element-wise product of two vectors $\boldsymbol{x} = (x_1, x_2, \dots x_d)^\top$ and $\boldsymbol{y} = (y_1, y_2, \dots y_d)^\top \in \mathbb{R}^d$ as $\boldsymbol{x} \odot \boldsymbol{y} = (x_1y_1, x_2y_2, \dots x_dy_d)$ and element-wise division by $\boldsymbol{x} \oslash \boldsymbol{y} = (\frac{x_1}{y_1}, \frac{x_2}{y_2} \dots \frac{x_d}{y_d})^\top$ provided $y_i \neq 0 \,\forall\, 1 \leq i \leq p$. For a matrix X, we denote its Frobenius norm by $\|X\|_F = \sqrt{\sum_{1\leq i\leq m} \sum_{1\leq j\leq d} x_{ij}^2}$. The minimum eigen value of X is represented by $\lambda_{\min}(X)$. We will sometimes use $(X)_{ii}$ instead of x_{ii} to represent the diagonal elements of matrix X. The symbol $\mathbb{R}^+$ denotes the strictly positive subset of the real numbers, i.e. $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x > 0\}$.

For a review of other useful mathematical notation used in our proofs see section 1.4, Table 3.1, and Table 3.2.

3.5.2.1 Lemmas

Lemma 3.5.2. *Let $A \in \mathbb{R}^{p\times p} \succ 0$ be symmetric positive definite and denote $\alpha > 0$ to be the minimum eigen value of A. Then we can write $A = \sum_{k=1}^{p} \lambda_k u_k u_k^\top + \alpha I_p$, where $\lambda_k \geq 0$ and $u_k \in \mathbb{R}^p$ are orthonormal for $1 \leq k \leq p$.*

This lemma basically says that, A is the sum of a positive semi-definite matrix and α times the identity matrix, where $\alpha > 0$ is the minimum eigen value of A.

Proof. We can write $A = (A - \alpha I_p) + \alpha I_p$ and we know that $A - \alpha I_p$ is symmetric (since A is symmetric).

Now if we can show that $A - \alpha I_p \succeq 0$ then we are done. This is because any real symmetric, positive semi-definite matrix is diagonalizable by orthogonal matrices and it has non-negative eigen values. That is, $A - \alpha I_p = U\Lambda U^\top = \sum_{k=1}^{p} \lambda_k u_k u_k^\top$, with $\lambda_k \geq 0$ for $1 \leq k \leq p$ and $UU^\top = U^\top U = I_p$.

By symmetry and positive definiteness of A we have, $A = \tilde{U}\tilde{\Lambda}\tilde{U}$ for diagonal $\tilde{\Lambda}$, with $(\tilde{\Lambda})_{ii} > 0$ for $1 \leq i \leq p$ and $\tilde{U}\tilde{U}^\top = \tilde{U}^\top\tilde{U} = I_p$. The diagonal entries $(\tilde{\Lambda})_{ii}$ are the eigen values of A.

We know that $A - \alpha I_p = \tilde{U}\tilde{\Lambda}\tilde{U}^\top - \alpha\tilde{U}\tilde{U}^\top = \tilde{U}(\tilde{\Lambda} - \alpha I_p)\tilde{U}^\top$ and $(\tilde{\Lambda})_{ii} - \alpha \geq 0$ (since $\alpha = \lambda_{\min}(A)$). Therefore $A - \alpha I_p$ is symmetric with non-negative eigen values, i.e $A - \alpha I_p \succeq 0$. We are done. □

Lemma 3.5.3. *For some integer $1 \leq K' \leq p-2$, let $A \in \mathbb{R}^{p\times p}$ be a matrix of the form $A = \sum_{k=1}^{K'+1} \zeta_k v_k v_k^\top + D$, where $\zeta_k \in \mathbb{R}^+$, $v_k \in \mathbb{R}^p$ are orthonormal for $1 \leq k \leq K'+1$ and $D \in \mathbb{R}^{p\times p}$ is a diagonal matrix with $(D)_{ii} > 0$. Then we can re-write A as:*

$$A = \sum_{k=1}^{K'+1} \tilde{\zeta}_k \tilde{v}_k \tilde{v}_k^\top + \tilde{D} \quad \text{where } \tilde{\zeta}_k \in \mathbb{R}^+,\ \tilde{v}_k \in \mathbb{R}^p \text{ are orthonormal.}$$

The matrix $\tilde{D} \in \mathbb{R}^{p\times p}$ is diagonal with $(\tilde{D})_{ii} = 0$ for $1 \leq i \leq K'+1$ and $(\tilde{D})_{ii} > 0$ for $K'+1 < i \leq p$.

Proof. We split the diagonal matrix D into two parts:

$$D = D_1 + D_2 = \begin{bmatrix} D_{1:K'+1,1:K'+1} & \boldsymbol{O}_1 \\ \boldsymbol{O}_1^\top & \boldsymbol{O}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{O}_3 & \boldsymbol{O}_1 \\ \boldsymbol{O}_1^\top & D_{K'+2:p,K'+2:p} \end{bmatrix}$$

Here, $\boldsymbol{O}_1 \in \mathbb{R}^{K'+1\times p-K'-1}$, $\boldsymbol{O}_2 \in \mathbb{R}^{p-K'-1\times p-K'-1}$ and $\boldsymbol{O}_3 \in \mathbb{R}^{K'+1\times K'+1}$ are matrices with all 0 entries.

We re-write A as:

$$A = (\underbrace{\sum_{k=1}^{K'+1} \zeta_k v_k v_k^\top + D_1}_{C}) + D_2$$

Now, since C and D_1 are symmetric positive semi-definite they are simultaneously diagonalizable [27]. Therefore, there exists an orthogonal matrix $\tilde{V} \in \mathbb{R}^{p\times p}$ and diagonal matrices $\Lambda_1, \Lambda_2 \in \mathbb{R}^{p\times p}$ such that:

$$C + D_1 = \tilde{V}\Lambda_1\tilde{V}^\top + \tilde{V}\Lambda_2\tilde{V}^\top = \tilde{V}(\Lambda_1 + \Lambda_1)\tilde{V}^\top$$

Since both C and D_1 are rank $K'+1$ matrices and positive semi-definite, we know that both Λ_1 and Λ_2 have exactly $K'+1$ strictly positive eigen values. Without loss of generality we assume

$(\Lambda_j)_{11} \geq (\Lambda_j)_{22} \ldots (\Lambda_j)_{(K'+1)(K'+1)} > 0$ for $j = 1, 2$.[7]

Therefore, we finally have: $C + D_1 = \sum_{k=1}^{K'+1} \big((\Lambda_1)_{kk} + (\Lambda_2)_{kk}\big)\tilde{v}_k\tilde{v}_k^\top = \sum_{k=1}^{K'+1} \tilde{\zeta}_k\tilde{v}_k\tilde{v}_k^\top$ for $\zeta_k = (\Lambda_1)_{kk} + (\Lambda_2)_{kk} > 0$ and orthonormal $\tilde{v}_k$ $1 \leq k \leq K' + 1$.

Observe that D_2 satisfies the conditions $(D_2)_{ii} = 0$ for $1 \leq i \leq K' + 1$ and $(D_2)_{ii} > 0$ for $K' + 1 < i \leq p$. We are done. □

3.5.2.2 Proofs of main result

For convenience of the reader, we re-state the main theorem below.

Theorem 3.5.4. *Consider the posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ defined in (3.40). Let $q_{\phi^*}^{IAF} \in Q^{IAF}$ be the optimal approximate posterior for $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$. Here, Q^{IAF} is the family generated by IAF with K hidden nodes in the shallow conditioner network c_ϕ defined in (3.4). That is:*

$$q_{\phi^*}^{IAF} \in argmin_{q_\phi^{IAF} \in Q^{IAF}} KL(q_\phi^{IAF} \,||\, \Pi(.|X, \boldsymbol{y}))$$

Let $\gamma_1 \geq \gamma_2 \cdots \geq \gamma_p$ be the eigen values of the covariance matrix $\Sigma_{\boldsymbol{\theta}}$. We assume $K = 2K'$ where $1 \leq K' \leq p - 2$. Then we have:

$$\min_{q_\phi^{IAF} \in Q^{IAF}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \sum_{k=K'+2}^{p} (\gamma_k - \gamma_p)^2 \tag{3.42}$$

Proof. We split this proof into two parts.

Step 1: Building on ideas in [2], we first bound the KL divergence for the family:

$$Q^{\text{SVI}} = \{N(m_\phi, \Sigma_\phi) \mid m_\phi \in \mathbb{R}^p,\ \Sigma_\phi \in \mathbb{R}^{p\times p} = \sum_{k=1}^{K'+1} \zeta_k v_k v_k^\top + \boldsymbol{D},\ \zeta_k > 0,\ v_k \in \mathbb{R}^p\}$$

The matrix $\boldsymbol{D}$ is a diagonal matrix with strictly positive diagonal entries $(D)_{ii} > 0$.[8]

[7] We can always re-order the eigen vectors $\tilde{v}_k$ to ensure this is satisfied.

[8] This is nothing but a rank $K' + 1$ plus diagonal structured variational family of Gaussians.

Step 2: Next we show that Q^{IAF} with $2K^{'}$ hidden nodes in c_ϕ contains the family defined above, i.e $Q^{\text{SVI}} \subseteq Q^{\text{IAF}}$.

We will then have:

$$\min_{q_\phi \in Q^{\text{IAF}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \sum_{k=K^{'}+2}^{p} (\gamma_k - \gamma_p)^2$$

Step 1:

We know that, the true posterior $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$ is distributed as $N(m_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}})$ where:

$$\Sigma_{\boldsymbol{\theta}} = (X^\top X + I_p)^{-1}$$

$$m_{\boldsymbol{\theta}} = {\Sigma_{\boldsymbol{\theta}}}^{-1} X^\top \boldsymbol{y}$$

Now observe that if q_ϕ is Gaussian $N(m_\phi, \Sigma_\phi)$, then we have:

$$KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) = \frac{1}{2}\Big\{ \text{Tr}({\Sigma_{\boldsymbol{\theta}}}^{-1}\Sigma_\phi) - p + \ln\big(\frac{\det \Sigma_{\boldsymbol{\theta}}}{\det \Sigma_\phi}\big) + (m_{\boldsymbol{\theta}} - m_\phi)^\top {\Sigma_{\boldsymbol{\theta}}}^{-1}(m_{\boldsymbol{\theta}} - m_\phi)\Big\}$$

Choosing $m_\phi = m_{\boldsymbol{\theta}}$ we have:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2}\Big\{ \text{Tr}({\Sigma_{\boldsymbol{\theta}}}^{-1}\Sigma_\phi - I_p) + \ln(\frac{\det \Sigma_{\boldsymbol{\theta}}}{\det \Sigma_\phi})\Big\}$$

We know, $KL(\Pi(.|X, \boldsymbol{y}) \,||\, q_\phi) \geq 0 \implies \ln(\frac{\det \Sigma_\phi}{\det \Sigma_{\boldsymbol{\theta}}}) \geq \text{Tr}(I_p - \Sigma_\phi^{-1}\Sigma_{\boldsymbol{\theta}})$

Continuing, we have:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2}\Big\{ \text{Tr}({\Sigma_{\boldsymbol{\theta}}}^{-1}\Sigma_\phi - I_p) - \ln(\frac{\det \Sigma_\phi}{\det \Sigma_{\boldsymbol{\theta}}})\Big\}$$

$$\leq \frac{1}{2}\,\text{Tr}({\Sigma_{\boldsymbol{\theta}}}^{-1}\Sigma_\phi + {\Sigma_\phi}^{-1}\Sigma_{\boldsymbol{\theta}} - 2I_p) = \frac{1}{2}\,\text{Tr}\left(\Sigma_\phi^{-1}(\Sigma_\phi - \Sigma_{\boldsymbol{\theta}})\Sigma_{\boldsymbol{\theta}}^{-1}(\Sigma_\phi - \Sigma_{\boldsymbol{\theta}})\right)$$

Applying Ruhe's trace inequality [37] for symmetric, positive semi-definite matrices we have:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2}\lambda_{\max}(\Sigma_\phi^{-1})\,\text{Tr}\left((\Sigma_\phi - \Sigma_{\boldsymbol{\theta}})\Sigma_{\boldsymbol{\theta}}^{-1}(\Sigma_\phi - \Sigma_{\boldsymbol{\theta}})\right)$$

$$\leq \frac{1}{2}\lambda_{\max}(\Sigma_\phi^{-1})\lambda_{\max}(\Sigma_{\boldsymbol{\theta}}^{-1})\|\Sigma_\phi - \Sigma_{\boldsymbol{\theta}}\|_F^2 = \frac{1}{2\lambda_{\min}(\Sigma_\phi)\lambda_{\min}(\Sigma_{\boldsymbol{\theta}})}\|\Sigma_\phi - \Sigma_{\boldsymbol{\theta}}\|_F^2$$

We note that $\Sigma_{\boldsymbol{\theta}} = (X^\top X + I_p)^{-1}$ is symmetric, positive definite. By Lemma 3.5.2 we can express it as: $\Sigma_{\boldsymbol{\theta}} = \sum_{k=1}^{p} \lambda_k u_k u_k^\top + \gamma_p I_p$, where $\gamma_p = \lambda_{\min}(\Sigma_{\boldsymbol{\theta}})$, $\lambda_k \geq 0$ and $u_k \in \mathbb{R}^p$ for $1 \leq k \leq p$ are

orthonormal. We then have the following:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \|\gamma_p I_p + \sum_{k=1}^{K'+1} \zeta_k v_k v_k^\top - \gamma_p I_p - \sum_{k=1}^{p} \lambda_k u_k u_k^\top\|_F^2$$

Without loss of generality let $\lambda_1 \geq \lambda_2 \cdots \geq \lambda_p$ and define $\tilde{U} \in \mathbb{R}^{p \times (p-K'-1)} = \left(u_{K'+2}, u_{K'+3}, \ldots u_p\right)$. We also define:

$$\tilde{\Lambda} = \begin{bmatrix} \lambda_{K'+2} & 0 & \cdots & 0 \\ 0 & \lambda_{K'+3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

Choosing $\zeta_k = \lambda_k$ and $v_k = u_k, 1 \leq k \leq K' + 1$, we get an upper bound on the KL as follows:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \| \sum_{k=K'+2}^{p} \lambda_k u_k u_k^\top\|_F^2$$

$$= \frac{1}{2\gamma_p^2} \text{Tr}\left((\tilde{U}\tilde{\Lambda}\tilde{U}^\top)(\tilde{U}\tilde{\Lambda}\tilde{U}^\top)^\top\right) = \frac{1}{2\gamma_p^2} \sum_{k=K'+2}^{p} \lambda_k^2$$

Now observe that if $U = (u_1, u_2, \ldots, \boldsymbol{u}_p)$ and $\Lambda \in \mathbb{R}^{p \times p}$ is a diagonal matrix with diagonal entires λ_k, $1 \leq k \leq p$, then the covariance matrix $\Sigma_{\boldsymbol{\theta}}$ can be written as:

$$\Sigma_{\boldsymbol{\theta}} = U\Lambda U^\top + \gamma_p U U^\top = U(\Lambda + \gamma_p I_p)U^\top \tag{3.43}$$

Therefore $\Sigma_{\boldsymbol{\theta}}$ has eigen values $\lambda_k + \gamma_p$ for $1 \leq k \leq p$. Note that since $\gamma_p = \lambda_{\min}(\Sigma_{\boldsymbol{\theta}})$, we have λ_p=0. Therefore, denoting the eigen values of $\Sigma_{\boldsymbol{\theta}}$ by γ_k we have:

$$\min_{q_\phi \in Q^{\text{SVI}}} KL(q_\phi \,||\, \Pi(.|X, \boldsymbol{y})) \leq \frac{1}{2\gamma_p^2} \sum_{k=K'+2}^{p} (\gamma_k - \gamma_p)^2$$

We now move onto the next step of the proof.

Step 2:

We now wish to show that $Q^{\text{SVI}} \subseteq Q^{\text{IAF}}$ where Q^{IAF} is the family generated by IAF with $2K'$ hidden nodes in c_ϕ. Q^{IAF} is generated by sampling $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots \theta_p^0) \sim N(0, I_p)$ and then computing $\boldsymbol{\theta} = \boldsymbol{a} \odot \boldsymbol{\theta}_0 + \boldsymbol{b}$ where $\boldsymbol{a} = (a_1, a_2, \ldots a_p)$ and $\boldsymbol{b} = (b_1, b_2, \ldots b_p)$. We present the detailed steps for sampling from $q_\phi \in Q^{\text{IAF}}$ below.

(i) Sample $\boldsymbol{\theta}_0 = (\theta_1^0, \theta_2^0, \ldots \theta_p^0) \sim N(0, I_p)$.

(ii) Compute $\theta_1 = a_1\theta_1^0 + b_1$ for some $a_1, b_1 \in \mathbb{R}$.

(iii) **FOR** $i = 2, 3 \ldots p$:

(iv) Compute $b_i\big(\boldsymbol{\theta}_{1:i-1}^0\big) = \sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=2j-1}^{\min\{2(i-1),2K'\}} v_{ik}^b g(c_{kj}^b\theta_j^0 + t_{\text{in},i}^b) + t_{\text{out},i}^b$

(v) Compute $a_i\big(\boldsymbol{\theta}_{1:i-1}^0\big) = h\Big(\sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=2j-1}^{\min\{2(i-1),2K'\}} v_{ik}^a g(c_{kj}^a\theta_j^0 + t_{\text{in},i}^a) + t_{\text{out},i}^a\Big)$

(vi) Now sample, $\theta_i = a_i\big(\boldsymbol{\theta}_{1:i-1}^0\big) \times \theta_i^0 + b_i\big(\boldsymbol{\theta}_{1:i-1}^0\big)$

(vii) **ENDFOR**

Above the functions $b_i\big(\boldsymbol{\theta}_{1:i-1}^0\big)$, $a_i\big(\boldsymbol{\theta}_{1:i-1}^0\big)$ follow from the definition of c_ϕ in (3.4). Here, $g(x) = \max(x, 0)$ is the ReLU function and $h : \mathbb{R} \to \mathbb{R}^+$ is invertible.

We will show $Q^{\text{SVI}} \subseteq Q_R^{\text{IAF}} \subseteq Q^{\text{IAF}}$, where Q_R^{IAF} is a restricted version of the IAF family of distributions satisfying:

$$
\begin{aligned}
&t_{\text{in},i}^a = t_{\text{in},i}^b = 0, \quad 1 \le i \le p \\
&c_{kj}^a = v_{ik}^a = 0 \;\forall\, i, j, k \implies a_i(\boldsymbol{\theta}_{1:i-1}^0) = h(t_{\text{out},i}^a) \\
&c_{(2k-1)j}^b = 1 \text{ and } c_{(2k)j}^b = -1, \quad j \le k \le \min\{i-1, K'\}, \;\; 1 \le j \le \min\{i-1, K'\} \\
&v_{i(2k-1)}^b = -v_{i(2k)}^b, \quad j \le k \le \min\{i-1, K'\}, \;\; 1 \le j \le \min\{i-1, K'\}
\end{aligned}
$$

Under the above constraints we can re-write $b_i(\boldsymbol{\theta}_{1:i-1}^0)$ as follows:

$$
\begin{aligned}
b_i(\boldsymbol{\theta}_{1:i-1}^0) &= \sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=j}^{\min\{i-1,K'\}} v_{i(2k-1)}^b g(c_{(2k-1)j}^b\theta_j^0) + v_{i(2k)}^b g(c_{(2k)j}^b\theta_j^0) + t_{\text{out},i}^b \\
&= \sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=j}^{\min\{i-1,K'\}} v_{i(2k-1)}^b g(\theta_j^0) - v_{i(2k-1)}^b g(-\theta_j^0) + t_{\text{out},i}^b \\
&= \sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=j}^{\min\{i-1,K'\}} v_{i(2k-1)}^b \theta_j^0 \mathbb{I}(\theta_j^0 \ge 0) - v_{i(2k-1)}^b(-\theta_j^0)\mathbb{I}(\theta_j^0 < 0) + t_{\text{out},i}^b \\
&= \sum_{j=1}^{\min\{i-1,K'\}} \sum_{k=j}^{\min\{i-1,K'\}} v_{i(2k-1)}^b \theta_j^0 + t_{\text{out},i}^b = \sum_{j=1}^{\min\{i-1,K'\}} w_{ij}^b \theta_j^0 + t_{\text{out},i}^b
\end{aligned}
$$

where $w^b_{ij} = \sum_{k=j}^{\min\{i-1,K'\}} v^b_{i(2k-1)}$.

Therefore, we finally have each transformed θ_i can be expressed as:

$$\theta_1 = a_1\theta_1^0 + b_1$$

$$\theta_i = a_i(\boldsymbol{\theta}^0_{1:i-1})\theta_i^0 + b_i(\boldsymbol{\theta}^0_{1:i-1}) = h(t^a_{\text{out},i})\theta_i^0 + \sum_{j=1}^{\min\{i-1,K'\}} w^b_{ij}\theta^0_j + t^b_{\text{out},i} \quad 2 \le i \le p \tag{3.44}$$

where $\theta_i^0 \overset{\text{i.i.d}}{\sim} N(0,1)$.

This can be equivalently written in matrix form as:

$$\boldsymbol{\theta} = \begin{bmatrix} a_1 & 0 & \dots & \dots & \dots & \dots & 0 \\ w^b_{21} & a_2 & 0 & \dots & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w^b_{K'1} & \dots & w^b_{K'(K'-1)} & a_{K'} & \dots & \dots & 0 \\ w^b_{(K'+1)1} & \dots & \dots & w^b_{(K'+1)K'} & a_{K'+1} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ w^b_{p1} & \dots & \dots & w_{pK'} & 0 & \dots & a_p \end{bmatrix} \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \\ \vdots \\ \theta_p^0 \end{bmatrix} + \begin{bmatrix} t^b_{\text{out},1} \\ t^b_{\text{out},2} \\ \vdots \\ t^b_{\text{out},p} \end{bmatrix}$$

$$= \left(\begin{bmatrix} B_1 & \boldsymbol{O}_1 \\ B_2 & \boldsymbol{O}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{O}_3 & \boldsymbol{O}_1 \\ \boldsymbol{O}_1^\top & \tilde{D} \end{bmatrix} \right) \begin{bmatrix} \theta_1^0 \\ \theta_2^0 \\ \vdots \\ \theta_p^0 \end{bmatrix} + \boldsymbol{t}^{\boldsymbol{b}}_{\text{out}}$$

where $B_1 \in \mathbb{R}^{(K'+1)\times(K'+1)} = \begin{bmatrix} a_1 & 0 & \dots & \dots & 0 \\ w^b_{21} & a_2 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w^b_{(K'+1)1} & w^b_{(K'+1)2} & \dots & w^b_{(K'+1)(K'-1)} & a_{K'+1} \end{bmatrix}$ is lower triangular.

$$B_2 \in \mathbb{R}^{(p-K'-1)\times(K'+1)} = \begin{bmatrix} w^b_{(K'+2)1} & \dots & w^b_{(K'+2)(K'+1)} \\ \vdots & \vdots & \vdots \\ w^b_{p1} & \dots & w_{p(K'+1)} \end{bmatrix}$$

$\tilde{D} \in \mathbb{R}^{(p-K^{'}-1)\times(p-K^{'}-1)} = \begin{bmatrix} a_{K^{'}+2} & 0 & \dots & 0 \\ 0 & a_{K^{'}+3} & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \dots & 0 & a_p \end{bmatrix}$ is a diagonal matrix and $\boldsymbol{O}_1 \in \mathbb{R}^{(K^{'}+1)\times(p-K^{'}-1)}$, $\boldsymbol{O}_2 \in \mathbb{R}^{(p-K^{'}-1)\times(p-K^{'}-1)}$, $\boldsymbol{O}_3 \in \mathbb{R}^{(K^{'}+1)\times(K^{'}+1)}$ are matrices with all 0 entries.

We use the following simplified notation:

$$L = \begin{bmatrix} B_1 & \boldsymbol{O}_1 \\ B_2 & \boldsymbol{O}_2 \end{bmatrix} \text{ and } D = \begin{bmatrix} \boldsymbol{O}_3 & \boldsymbol{O}_1 \\ \boldsymbol{O}_1^\top & \tilde{D} \end{bmatrix} \tag{3.45}$$

Observe that w_{ij}^b and a_i can be selected such that $L \in \mathbb{R}^{p\times p}$ is any lower triangular matrix with $(K^{'}+1)$ strictly positive diagonal elements and remaining $p-K^{'}-1$ columns containing all zeros. Further, D^2 is any diagonal matrix with $p-K^{'}-1$ strictly positive lower diagonal elements.

Since $\boldsymbol{\theta}_0 \sim N(0, I_p)$, we have: $\boldsymbol{\theta} \sim N(\boldsymbol{t}_{\text{out}}^{\boldsymbol{b}}, (L+D)(L+D)^\top)$.

It is easily verified that $(L+D)(L+D)^\top$ simplifies to $LL^\top + D^2$. Therefore:

$$Q_R^{\text{IAF}} = \{N(\boldsymbol{t}_{\text{out}}^{\boldsymbol{b}}, \Sigma_\phi) \mid \boldsymbol{t}_{\text{out}}^{\boldsymbol{b}} \in \mathbb{R}^p,\ \Sigma_\phi = LL^\top + \overline{D}\} \subseteq Q^{\text{IAF}} \tag{3.46}$$

where L and $\overline{D}$ satisfy (3.45). Here we have used $\overline{D} = D^2$.

Now we will connect Q_R^{IAF} to Q^{SVI}. We recollect that Q^{SVI} is a family of gaussians with mean $m_\phi \in \mathbb{R}^p$ and covariance matrix $\Sigma_\phi \in \mathbb{R}^{p\times p} = \sum_{k=1}^{K^{'}+1} \zeta_k v_k v_k^\top + \boldsymbol{D},\ \zeta_k > 0$ and $v_k \in \mathbb{R}^p$. That is, Σ_ϕ is the sum of a rank $K^{'}+1$ matrix and a diagonal matrix with strictly positive entries.

By Lemma 3.5.3 we know that Σ_ϕ can be re-written as:

$$\Sigma_\phi = \sum_{k=1}^{K^{'}+1} \tilde{\zeta}_k \tilde{v}_k \tilde{v}_k^\top + \begin{bmatrix} \boldsymbol{O}_3 & \boldsymbol{O}_1 \\ \boldsymbol{O}_1^\top & \boldsymbol{D}_{K^{'}+2:p,\ K^{'}+2:p} \end{bmatrix}$$

Now observe that $C = \sum_{k=1}^{K'+1} \tilde{\zeta}_k \tilde{v}_k \tilde{v}_k^\top$ is a rank $K' + 1$, symmetric positive semi-definite matrix. Therefore it can be expressed as $LL^\top$, for some L which is lower triangular with $K' + 1$ positive entries on the diagonal and remaining $p - K' - 1$ columns containing only zeros [11].

Therefore, Q^{SVI} is of the form:

$$Q^{\text{SVI}} = \{N(m_\phi, \Sigma_\phi) \mid m_\phi \in \mathbb{R}^p,\ \Sigma_\phi \in \mathbb{R}^{p \times p} = LL^\top + D\} \tag{3.47}$$

For lower triangular L and diagonal D as per (3.45). Since both m_ϕ and $\boldsymbol{t}_{\text{out}}^{\boldsymbol{b}}$ are allowed to vary freely in $\mathbb{R}^p$ by definitions in (3.46) and (3.47) we have $Q^{\text{SVI}} \subseteq Q_R^{\text{IAF}} \subseteq Q^{\text{IAF}}$. We are done. □

CHAPTER 4

LINEAR REGRESSION WITH SPIKE AND SLAB PRIORS

In many scientific problems, we wish to find out which variables are associated with a particular response. Statistics literature terms this the variable selection problem and there is a multitude of research in this direction. One of the simplest and most common formulations for this problem is a linear regression model of the form:

$$\boldsymbol{y} = X\boldsymbol{\theta} + \epsilon \tag{4.1}$$

where $\boldsymbol{y} = (y_1, y_2 \dots y_n)$, $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots \boldsymbol{x}_p) \in \mathbb{R}^{n \times p}$ and $\epsilon \in \mathbb{R}^n \sim N(0, \sigma^2 I_n)$ for some $\sigma^2 > 0$. The goal is to select the variables (predictors) that are significantly associated with the response.

In recent years, a major stream of research has surfaced wherein we consider this problem under the high-dimensional regime $n \leq p$ (usually $n << p$). In order to conduct statistical inference in such a set up we require sparsity constraints on the underlying true model, that is, there exists some set $\mathcal{M} = \{i \,|1 \leq i \leq p \text{ and } \theta_i \neq 0\}$ such that the cardinality of $\mathcal{M}$ satisfies $|\mathcal{M}| < n$. These sparsity constraints impose the assumption that, despite the large number of variables under consideration, the set of variables that are actually useful for predicting the response has a cardinality smaller than the sample size.

There are many ways to deal with variable selection in high-dimension. Perhaps the most well known of these is the Lasso [40], which introduces an L_1 penalty term for $\boldsymbol{\theta}$ into the sum of squared errors loss function. Despite exhibiting desirable properties such as selection consistency and computational efficiency, the lasso and related penalized regression methods run into issues when multiple correlated variables occur in the design matrix X. More specifically, in a group of multiple correlated predictors the lasso tends to select only one of these predictors and sets the others to zero rather than spreading out the contribution of this group between individual variables. This leads to a loss of information [13].

With the advent of powerful computational resources, Bayesian approaches to variable selection have gained traction in the statistics community. In Bayesian variable selection (BVS), sparsity in the model parameters is induced via the prior distribution $\pi(\boldsymbol{\theta})$. Some of the advantages of

BVS include uncertainty quantification for our estimates of $\boldsymbol{\theta}$, as well as the calculation of easily interpretable posterior inclusion probabilities $\mathbb{P}(\theta_i \neq 0)$. The paper [13] has simulation studies showing that BVS manages to spread out the posterior inclusion probabilies among the multiple correlated predictors in a group as opposed to the lasso, thus avoiding the loss of information mentioned earlier.

Among the wide variety of ways in which we can select the prior $\pi(\boldsymbol{\theta})$, one of the most enduring approaches is the spike and slab prior. Spike and slab prior distributions take the following form:

$$\theta_i \overset{\text{i.i.d}}{\sim} w f_1(\boldsymbol{\theta}) + (1 - w) f_2(\boldsymbol{\theta})$$

where $0 < w < 1$. Observe that each θ_i is a mixture of 2 probability distributions f_1 and f_2 with weights w and $1 - w$ respectively. The common convention is to choose the distribution $f_1(.)$ to be any continuous distribution on $\mathbb{R}^p$ with a non-negligible variance (the slab).[1] The distribution f_2 is chosen to be a distribution with very small or zero variance (the spike). Often, a prior distribution is imposed on w as well, e.g. $w \sim \text{Beta}(a_0, b_0)$.

The main challenge in implementing BVS with spike and slab priors is accurately and efficiently sampling from the posterior distribution. Sampling from the true posterior requires marginalizing over 2^p possible models and is computationally infeasible even for moderate dimensions. MCMC methods while statistically accurate, may have a very slow mixing time (running for days to produce samples close to those from the target distribution) due to high multimodality of the true posterior as p increases. Mean-Field VI (MF-VI) has gained immense popularity for sampling from intractable posteriors resulting from using spike and slab BVS. The paper [31] has used MF-VI and Structured VI (SVI) for linear regression with a spike and slab prior and includes derivations for oracle contraction rates of the variational posterior. [7] and [41] have used algorithms based on MF-VI for genetic association studies, with a reasonable degree of success. [23] covers the case of MF-VI for grouped linear regression, in which the predictors have an underlying group structure.

As dicussed throughout this dissertation, MF-VI has certain disadvantages when there are multiple groups of correlated predictor variables. The papers [7] and [31] show via simulations

[1] The slab is generally selected to have tails that are atleast as heavy as exponential tails for better performance [31].

that the variational posterior from MF-VI under-estimates the posterior variance and is therefore not very reliable for uncertainty quantification. This is consistent with our theoretical results in chapter 3.

In this chapter we will explore the use of Normalizing Flows aided Variational Inference (FAVI) to improve uncertainty quantification in BVS with spike and slab priors. Keeping with the themes of the rest of this dissertation, we focus on variable selection with the linear regression formulation in (4.1). The main challenge in adapting FAVI to spike and slab regression is that we need to jointly model the continuous $\boldsymbol{\theta}$ and discrete latent variables $\boldsymbol{z} \in \{0,1\}^p$. We will elaborate on this in section 4.0.2.

4.0.1 Model

We assume the linear model defined in (4.1). We will later discuss adaptations to the generalized linear model (glm) in section 4.4.

We adopt a standard spike and slab formulation for the prior distribution $\pi(\boldsymbol{\theta})$, similar to works [31] and [32] as follows:

$$\theta_i \overset{\text{i.i.d}}{\sim} w f_\pi + (1-w)\delta_0, \quad w \in (0,1), \quad 1 \leq i \leq p$$

In the equation above, f_π could be any continuous distribution with non-zero support on $\mathbb{R}^p$. δ_0 denotes the dirac delta function (point mass at 0). For any Borel set $A \in B(\mathbb{R})$ we have:

$$\delta_0(A) = \begin{cases} 1, & \text{if } 0 \in A \\ 0, & \text{otherwise} \end{cases}$$

Since we have no additional information to suggest otherwise, we assume $w = \frac{1}{2}$. To allow for easier calculations later on we re-write the prior in the following manner:

$$\begin{aligned} \theta_i &\overset{i.i.d}{\sim} z_i f_\pi + (1-z_i)\delta_0 \\ z_i &\overset{i.i.d}{\sim} \text{Bernoulli}(w), \quad 1 \leq i \leq p \end{aligned}$$

Our aim is to sample from the posterior distribution $\Pi(\boldsymbol{\theta}, \boldsymbol{z}|X, \boldsymbol{y})$ given in (4.2).[2]

$$\begin{aligned}\Pi(\boldsymbol{\theta}, \boldsymbol{z}|X, \boldsymbol{y}) &= \frac{\mathbb{P}(\boldsymbol{y}|X, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{z})\pi(\boldsymbol{z})}{\int_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} \mathbb{P}(\boldsymbol{y}|X, \boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{z})\pi(\boldsymbol{z})} \\ &= \frac{(\sqrt{2\pi}\sigma)^{-n} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\theta}\|_2^2} \prod_{i=1}^{p} \left\{ \big(f_\pi(\theta_i)\big)^{z_i} \big(\delta_0(\theta_i)\big)^{1-z_i} w^{z_i}(1-w)^{z_i} \right\}}{\int_{\boldsymbol{\theta}} \sum_{\boldsymbol{z}} (\sqrt{2\pi}\sigma)^{-n} e^{-\frac{1}{2\sigma^2}\|\boldsymbol{y}-X\boldsymbol{\theta}\|_2^2} \prod_{i=1}^{p} \left\{ \big(f_\pi(\theta_i)\big)^{z_i} \big(\delta_0(\theta_i)\big)^{1-z_i} w^{z_i}(1-w)^{z_i} \right\}}\end{aligned} \tag{4.2}$$

A key advantage of assuming a dirac spike δ_0 is that it facilitates the direct calculation of the posterior inclusion probabilites $\mathbb{P}(\theta_i \neq 0)$ without needing to run any additional hypothesis tests. It also improves computational efficiency for approximate inference methods by ensuring a large number of $\theta_i = 0$. For instance, in each iteration of Gibbs sampling we will only need to work with the predictors $\boldsymbol{x}_i$ for $1 \leq i \leq p$ such that $\theta_i \neq 0$. Under a reasonable choice of hyper-parameters for the prior and if the sparsity assumptions hold true, the cardinality of the set $\mathcal{M} = \{i \mid 1 \leq i \leq p, \text{ and } \theta_i \neq 0\}$ will be small ($|\mathcal{M}| << \min\{n, p\}$).

4.0.2 Method

We introduce an algorithm leveraging FAVI to sample from the target posterior $\Pi(\boldsymbol{\theta}, \boldsymbol{z}|X, \boldsymbol{y})$ defined in (4.2). This method is inspired by the work in [45]. The idea is to condition the parameters $\boldsymbol{\theta} \in \mathbb{R}^p$ which have a continuous distribution on the discrete latent variables, $\boldsymbol{z} \in \mathbb{R}^p$ and use normalizing flows only for $\boldsymbol{\theta}$. We assume a mean-field family for the discrete component $\boldsymbol{z} \in \mathbb{R}^p$. By assuming a mean-field family for $\boldsymbol{z}$ we transform the problem of modelling a joint discrete distribution with $O(2^p)$ parameters to a distribution with only $O(p)$ variational parameters.

4.1 Variational Family

We define the variational family of distributions below.

$$\begin{aligned}q_{\phi,\psi,\boldsymbol{r}}(\boldsymbol{\theta}, \boldsymbol{z}) &= q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z}) q_{\boldsymbol{r}}(\boldsymbol{z}) \\ q_{\boldsymbol{r}}(\boldsymbol{z}) &= \prod_{i=1}^{p} r_i^{z_i}(1-r_i)^{1-z_i}\big(\mathbb{I}\{z_i = 0\} + \mathbb{I}\{z_i = 1\}\big), \quad 0 < r_i < 1\end{aligned}$$

Above we have assumed $\boldsymbol{z}_i \overset{\text{i.i.d}}{\sim} \text{Bernoulli}(r_i)$. To model $q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})$ we take the following approach:

[2]Note that if we have samples $(\boldsymbol{\theta}^{(l)}, \boldsymbol{z}^{(l)})$, $1 \leq l \leq L$ from $\Pi(\boldsymbol{\theta}, \boldsymbol{z}|X, \boldsymbol{y})$ then $\boldsymbol{\theta}^{(l)}$ are samples from $\Pi(\boldsymbol{\theta}|X, \boldsymbol{y})$.

(i) Sample $\boldsymbol{\theta}_0$ from a mean-field spike and slab base distribution

$$q_\psi(\boldsymbol{\theta}_0|\boldsymbol{z}) = \prod_{i=1}^{p} q_{\psi_i}(\theta_i^0) = \prod_{i=1}^{p} \big(f_{\psi_i}(\theta_i)\big)^{z_i} \big(\delta_0(\theta_i)\big)^{1-z_i}$$

The distribution f_{ψ_i} belongs to the same family as f_π (the slab component in the prior).

(ii) Apply a normalizing flow transformation T_ϕ to the samples $\boldsymbol{\theta}_0$ from the base distribution to get $\boldsymbol{\theta}_{S+1} = \boldsymbol{z} \odot (T_S \circ T_{S-1} \cdots \circ T_1(\boldsymbol{\theta}_0))$, where $\odot$ is the element-wise product between two vectors.

We use Neural auto-regressive flows (NAF) with the Deep Sigmoidal Flow (DSF) for T_ϕ. We review the definition of the NAF transformation below.

$$\theta_i^s = \sigma^{-1}(\boldsymbol{w}_i^\top \sigma(\boldsymbol{a}_i \theta_i^{s-1} + \boldsymbol{b}_i)), \quad 1 \le i \le p,\ 1 \le s \le S$$

The DSF parameters $\boldsymbol{w}_i,\ \boldsymbol{a}_i,\ \boldsymbol{b}_i \in \mathbb{R}^H$ are the outputs of conditioner networks $c_\phi^{\boldsymbol{w}}(.),\ c_\phi^{\boldsymbol{b}}(.),\ c_\phi^{\boldsymbol{a}}(.) : \mathbb{R}^p \to \mathbb{R}^{pH}$. Further, $\boldsymbol{a}_i$ and $\boldsymbol{w}_i$ are constrained as $a_{i,j} > 0\ \forall i, j, 0 < w_{i,j} < 1, \sum_j w_{i,j} = 1$ to ensure the invertibility of the DSF transformation.

4.2 Implementation details

In order to implement FAVI we need to maximize the Evidence Lower Bound (ELBO) between the variational family and the target posterior (see (2.2) for a review):

$$\begin{aligned} ELBO(q_{\phi,\psi,\boldsymbol{r}} \,||\, \Pi(.|X,\boldsymbol{y})) &= \mathbb{E}_{q_\phi(\boldsymbol{\theta})}\big[\ln\big(p(\boldsymbol{y}|X,\boldsymbol{\theta})\pi(\boldsymbol{\theta},\boldsymbol{z})\big)\big] - \mathbb{E}_{q_{\phi,\psi,\boldsymbol{r}}(\boldsymbol{\theta},\boldsymbol{z})}\big[\ln q_{\phi,\psi,\boldsymbol{r}}(\boldsymbol{\theta},\boldsymbol{z})\big] \\ &= \mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})q_{\boldsymbol{r}}(\boldsymbol{z})}[L(\boldsymbol{\theta})] - KL(q_{\phi,\psi,\boldsymbol{r}} \,||\, \pi(.)) \end{aligned}$$

where $L(\boldsymbol{\theta}) = \ln p(\boldsymbol{y}|X,\boldsymbol{\theta})$. Let $I = \{1 \le i \le p | z_i = 1\}$ and $I^c = \{1 \le i \le p | z_i = 0\}$ and assume $\sigma = 1$. This assumption is consistent with related work [31]. If we expect that $\sigma \neq 1$, then we can obtain an estimate $\hat{\sigma}$ of σ from the data and use the model $\hat{\sigma}^{-1}\boldsymbol{y} = \hat{\sigma}^{-1}X\boldsymbol{\theta} + \hat{\sigma}^{-1}\epsilon$. Let $X_I = (x_{i_1}, x_{i_2}, \ldots x_{i_{|I|}})$, where $i_k \in I$ for $1 \le k \le |I|$; that is, X_I is the submatrix of selected predictors. Similarly, let $\boldsymbol{\theta}_I = (\theta_{i_1}, \theta_{i_2}, \ldots \theta_{i_{|I|}})$.

We have:

$$\begin{aligned}
&\textbf{ELBO}(q_{\phi,\psi,\boldsymbol{r}} \,||\, \Pi(.|X,\boldsymbol{y})) = \mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})q_{\boldsymbol{r}}(\boldsymbol{z})}[L(\boldsymbol{\theta})] - KL(q_{\phi,\psi,\boldsymbol{r}} \,||\, \pi(.)) \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(-\frac{1}{2}\|\boldsymbol{y} - X_I\boldsymbol{\theta}_I^S\|_2^2 - \frac{1}{2}\ln 2\pi\Big)\right]}_{I} - \underbrace{\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})}\Big(\ln\frac{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})}{\pi(\boldsymbol{\theta}|\boldsymbol{z})}\Big)\right]}_{II} \\
&- \underbrace{KL(q_{\boldsymbol{r}}(\boldsymbol{z}) \,||\, \pi(\boldsymbol{z}))}_{III}
\end{aligned} \tag{4.3}$$

We will now simplify the ELBO in order to more easily implement stochastic gradient ascent (SGA).

Analyzing Term I: Due to the complex structure of T_ϕ we have to resort to MC sampling for this term and we do not simplify further.

Analyzing Term II:

$$\begin{aligned}
&\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})}\Big(\ln\frac{q_{\phi,\psi}(\boldsymbol{\theta}|\boldsymbol{z})}{\pi(\boldsymbol{\theta}|\boldsymbol{z})}\Big)\right] = \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{I^c},\boldsymbol{z})\delta_{\boldsymbol{0}}(\boldsymbol{\theta}_{I^c})}\Big(\ln\frac{q_{\phi,\psi}(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{I^c},\boldsymbol{z})\delta_{\boldsymbol{0}}(\boldsymbol{\theta}_{I^c})}{\pi(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{I^c},\boldsymbol{z})\delta_{\boldsymbol{0}}(\boldsymbol{\theta}_{I^c})}\Big)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{q_{\phi,\psi}(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{I^c}=0,\boldsymbol{z})}\Big(\ln\frac{q_{\phi,\psi}(\boldsymbol{\theta}_I|\boldsymbol{\theta}_{I^c}=0,\boldsymbol{z})}{\prod_{i\in I} f_\pi(\theta_i)}\Big)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\left(\ln\frac{\prod_{i\in I} f_{\psi_i}(\theta_i^0)}{\prod_{i\in I} f_\pi(\theta_i^S)} - \sum_{s=1}^{S}\sum_{i\in I}\ln\Big|\det\Big(\frac{\partial T_s}{\partial\theta_i^{s-1}}\Big)\Big|\right)\right] \\
&= \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\sum_{i\in I}\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\ln\frac{f_{\psi_i}(\theta_i^0)}{f_\pi(\theta_i^S)}\Big)\right] - \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\sum_{i\in I}\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\sum_{s=1}^{S}\ln\Big|\det\Big(\frac{\partial T_s}{\partial\theta_i^{s-1}}\Big)\Big|\Big)\right] \\
&= \sum_{i=1}^{p}\left\{\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\ln\frac{f_{\psi_i}(\theta_i^0)}{f_\pi(\theta_i^S)}\Big)\right] - \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\sum_{s=1}^{S}\ln\Big|\det\Big(\frac{\partial T_s}{\partial\theta_i^{s-1}}\Big)\Big|\Big)\right]\right\}
\end{aligned}$$

Simplifying further we have:

$$\begin{aligned}
II &= \sum_{i=1}^{p}\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\left\{\mathbb{E}_{f_{\psi_i}(\theta_i^0)}\Big(\ln f_{\psi_i}(\theta_i^0)\Big) - \mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\ln f_\pi(\theta_i^S)\Big)\right\}\right] \\
&- \sum_{i=1}^{p}E_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\sum_{s=1}^{S}\ln\Big|\det\Big(\frac{\partial T_s}{\partial\theta_i^{s-1}}\Big)\Big|\Big)\right] \\
&= \sum_{i=1}^{p}\left\{r_i\mathbb{E}_{f_{\psi_i}(\theta_i^0)}\Big(\ln f_{\psi_i}(\theta_i^0)\Big) - \mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\ln f_\pi(\theta_i^S)\Big)\right]\right. \\
&\left. - E_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\sum_{s=1}^{S}\ln\Big|\det\Big(\frac{\partial T_s}{\partial\theta_i^{s-1}}\Big)\Big|\Big)\right]\right\}
\end{aligned}$$

Analyzing Term III:

$$KL(q_{\boldsymbol{r}}(\boldsymbol{z})||\pi(\boldsymbol{z})) = \sum_{i=1}^{p} \left\{ r_i \ln \frac{r_i}{w} + (1-r_i) \ln \frac{1-r_i}{1-w} \right\} \tag{4.4}$$

Therefore we finally have the ELBO equal to:

$$\begin{aligned}
&\textbf{ELBO}(q_{\phi,\psi,\boldsymbol{r}} \,||\, \Pi(.|X,\boldsymbol{y})) \\
&= \underbrace{\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[\mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(-\frac{1}{2}\|\boldsymbol{y} - X_I \boldsymbol{\theta}_I^S\|_2^2 - \frac{1}{2}\ln 2\pi \Big) \right]}_{I} - \sum_{i=1}^{p} \left\{ r_i \ln \frac{r_i}{w} + (1-r_i) \ln \frac{1-r_i}{1-w} \right\} \\
&- \underbrace{\sum_{i=1}^{p} r_i \mathbb{E}_{f_{\psi_i}(\theta_i^0)}\Big(\ln f_{\psi_i}(\theta_i^0) \Big)}_{II(a)} + \sum_{i=1}^{p} \underbrace{\mathbb{E}_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i \mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\Big(\ln f_{\pi}(\theta_i^S) \Big) \right]}_{II(b)} \\
&+ \sum_{i=1}^{p} \underbrace{E_{q_{\boldsymbol{r}}(\boldsymbol{z})}\left[z_i \mathbb{E}_{(\prod_{i\in I} f_{\psi_i}(\theta_i^0))}\left(\sum_{s=1}^{S-1} \ln \Big| \det\Big(\frac{\partial T_s}{\partial \theta_i^{s-1}} \Big) \Big| \right) \right]}_{II(c)}
\end{aligned}$$

We use MC samples for terms I, $II(b)$ and $II(c)$, but $II(a)$ can usually be solved explicitly depending on our choice of f_ψ.

The re-parameterization trick: Denote $\textbf{ELBO}(q_{\phi,\psi,\boldsymbol{r}} \,||\, \Pi(.|X,\boldsymbol{y}))$ by $\mathcal{L}(q_{\phi,\psi,\boldsymbol{r}})$. We wish to maximize $\mathcal{L}(q_{\phi,\psi,\boldsymbol{r}})$ (minimize $-\mathcal{L}(q_{\phi,\psi,\boldsymbol{r}})$) with respect to $\phi, \psi, \boldsymbol{r}$. In stochastic gradient ascent, we calculate $l(q_{\phi,\psi,\boldsymbol{r}})$, which is a realization for an unbiased estimator of $\nabla_{\phi,\psi,\boldsymbol{r}} \mathcal{L}(q_{\phi,\psi,\boldsymbol{r}})$ (see 2.5.1). In order to calculate $l(q_{\phi,\psi,\boldsymbol{r}})$, we need to compute functions of the form $\frac{\partial}{\partial\gamma}\mathbb{E}_{u\sim q_\gamma}[f(u)]$, where q_γ is a probability distribution.[3] If we can write $u = \tilde{g}(\gamma, \epsilon)$ for a function $\tilde{g}(.) : \mathbb{R} \to \mathbb{R}$ and random variable $\epsilon \sim p_\epsilon$ that is free of the parameter γ, then: $\frac{\partial}{\partial\gamma}\mathbb{E}_{u\sim q_\gamma}[f(u)] = \mathbb{E}_{p_\epsilon}[\frac{\partial f}{\partial \tilde{g}}\frac{\partial \tilde{g}}{\partial \gamma}]$. For $u \sim N(m, s^2)$, we use $u = s\epsilon + m$ and $\epsilon \sim N(0,1)$. For $u \sim \text{Bernoulli}(r_i)$, we use the Gumbel softmax transformation $u = 1/(1+e^{-(-\ln(-\ln\epsilon)+\tilde{r}_i)/\lambda_0})$; where $\epsilon \sim \text{Uniform}(0,1)$, λ_0 is a temperature parameter and $\tilde{r}_i = \ln(r_i/(1-r_i))$. As $\lambda_0 \to 0$, the pdf of u converges to a $\text{Bernoulli}(r_i)$ distribution ([17], [25]). In practice we set $z_i = \mathbb{I}\{u > 0.5\}$ during the forward pass but when differentiating with respect to the variational parameters in the backward pass we use the continuous relaxation u (this is known as the Gumbel-softmax Straight-Through (ST) transformation).

[3]Note that in this case u corresponds to θ_i or $\boldsymbol{z}_i$ and q_γ is $N(m_i, s_i)$ for θ_i and $\text{Bernoulli}(r_i)$ for z_i respectively.

Algorithm 4.1 Normalizing **F**lows **A**ided **V**ariational **I**nference for **s**pike and **s**lab **reg**ression (**FAVI_ssreg**)

Inputs: training data $D = (X, \boldsymbol{y})$, number of flow transformations S, softmax temperature λ_0, learning rate α_0, exponential decay rates for momentum in Adam optimizer β_0, β_1, number of flow samples L, batchsize for training data $bsize$ and number of batches $nbatch$, number of layers L_c and hidden dimension d_c for cMADE conditioner networks $c_\phi(.)$, hidden dimension H for DSF flows.

Method:

Set initial values of parameters $\tilde{\boldsymbol{r}}$, $\psi = (m_i, s_i)_{i=1}^p$ and ϕ.

while not converged **do**

for $i = 1, 2 \ldots, nbatch$ **do**

Sample mini-batch D_{mini} from data D.

for $l = 1, 2, \ldots L$ **do**

Sample $\tilde{z}_i^{(l)} \overset{\text{i.i.d}}{\sim}$ Gumbel-softmax-ST(r_i, λ_0), $r_i = \sigma(\tilde{r}_i) = 1/(1 + e^{-\tilde{r}_i})$ for $1 \le i \le p$.

If $z_i^{(l)} = 0$ set $\theta_i^{0(l)} = 0$. If $z_i^{(l)} = 1$, independently sample $\theta_i^{0(l)} \sim N(m_i, s_i)$.

end for

for $s = 1, 2 \ldots S$ **do**

Apply c_ϕ to get $(\boldsymbol{w}_i^{s(l)}, \boldsymbol{b}_i^{s(l)}, \boldsymbol{a}_i^{s(l)}) = c_\phi(\boldsymbol{\theta}_{1:i-1}^{s-1(l)})$, $1 \le l \le L$ and $1 \le i \le p$.

Compute $\theta_i^{s(l)} = z_i^{(l)} \sigma^{-1}(\boldsymbol{w}_i^{s(l)\top} \sigma(\boldsymbol{a}_i^{s(l)} \theta_i^{s-1(l)} + \boldsymbol{b}_i^{s(l)}))$, $\quad 1 \le i \le p,\ 1 \le l \le L$

end for

Calculate unbiased estimate of **ELBO** as:

$$
\begin{aligned}
l(\phi, \psi, \tilde{\boldsymbol{r}}) = \frac{1}{L} \sum_{l=1}^{L} \Bigg\{ & \sum_{i=1}^{bsize} \Big[-\frac{1}{2} \sum_{i=1}^{bsize} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}^{S+1(l)})^2 - \frac{n}{2} \ln 2\pi \Big] \\
& - \frac{1}{nbatch} \sum_{i=1}^{p} \Big[\sigma(\tilde{r}_i) \ln \frac{\sigma(\tilde{r}_i)}{w} + (1 - \sigma(\tilde{r}_i)) \ln \frac{1 - \sigma(\tilde{r}_i)}{1 - w} \Big] \\
& - \frac{1}{nbatch} \sum_{i=1}^{p} \sigma(\tilde{r}_i) \mathbb{E}_{f_{\psi_i}(\theta_i^0)} \Big(\ln f_{\psi_i}(\theta_i^0) \Big) + \frac{1}{nbatch} \sum_{i=1}^{p} z_i^{(l)} \ln f_\pi(\theta_i^{S(l)}) \\
& + \frac{1}{nbatch} \sum_{i=1}^{p} z_i \sum_{s=1}^{S} \ln \Big| \det \Big(\frac{\partial T_s}{\partial \theta_i^{s-1(l)}} \Big) \Big| \Bigg\}
\end{aligned}
$$

Update ψ as $\psi = \psi + \alpha_0 \nabla_\psi l(\phi, \psi, \tilde{\boldsymbol{r}})$

Update ϕ as $\phi = \phi + \alpha_0 \nabla_\phi l(\phi, \psi, \tilde{\boldsymbol{r}})$

Update $\tilde{\boldsymbol{r}}$ as $\tilde{\boldsymbol{r}} = \tilde{\boldsymbol{r}} + \alpha_0 \nabla_{\tilde{\boldsymbol{r}}} l(\phi, \psi, \tilde{\boldsymbol{r}})$

end for

end while

Output: Converged $\phi^*, \psi^*, \tilde{\boldsymbol{r}}^*$

4.3 Simulation study

For our simulation study we assume a Gaussian prior for the slab f_π and a Gaussian variational distribution as follows:

$$\textbf{Prior: } f_\pi(\theta_i) = \frac{1}{\sqrt{2\pi}\tau} e^{-\frac{\theta_i^2}{2\tau^2}} \quad \textbf{Variational distribution: } f_{\psi_i}(\theta_i) = \frac{1}{\sqrt{2\pi}s_i} e^{-\frac{(\theta_i - m_i)^2}{2s_i^2}} \tag{4.5}$$

The unbiased estimator of the ELBO then simplifies to:

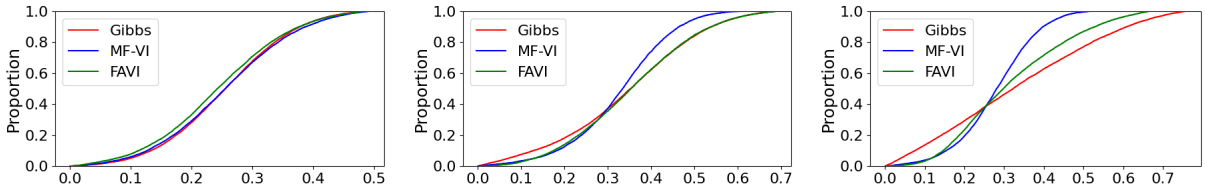
$$\begin{aligned} l(\phi, \psi, \tilde{\boldsymbol{r}}) = \frac{1}{L}\sum_{l=1}^{L}\Bigg\{ & -\frac{1}{2}\sum_{i=1}^{bsize}(y_i - \boldsymbol{x}_i^\top \boldsymbol{\theta}^{S(l)})^2 - \frac{n}{2}\ln 2\pi \\ & - \frac{1}{nbatch}\sum_{i=1}^{p}\Big[\sigma(\tilde{r}_i)\ln\frac{\sigma(\tilde{r}_i)}{w} + (1-\sigma(\tilde{r}_i))\ln\frac{1-\sigma(\tilde{r}_i)}{1-w}\Big] \\ & - \frac{1}{nbatch}\sum_{i=1}^{p}\sigma(\tilde{r}_i)\Big(-\frac{1}{2} - \frac{\ln 2\pi}{2} - \ln s_i\Big) + \frac{1}{nbatch}\sum_{i=1}^{p} z_i^{(l)}\Big(-\ln\tau - \frac{\ln 2\pi}{2} - \frac{(\theta_i^{S(l)})^2}{2\tau^2}\Big) \\ & + \frac{1}{nbatch}\sum_{i=1}^{p} z_i^{(l)}\sum_{s=1}^{S}\ln\Big|\det\Big(\frac{\partial T_s}{\partial \theta_i^{s-1(l)}}\Big)\Big|\Bigg\} \end{aligned}$$

The simulation set-up we use is similar to that of section 2.6.3.1. We assume $\tau = 1$ and the p-dimensional predictor variables $\boldsymbol{x}_1, \boldsymbol{x}_2 \ldots \boldsymbol{x}_n \in \mathbb{R}^p$ are simulated from a multivariate normal distribution $N(0, \Sigma)$. The covariance matrix $\Sigma = (1-\rho)I_p + \rho\mathbf{1}\mathbf{1}^\top$, where $\mathbf{1} = (1, 1, \ldots 1)^\top \in \mathbb{R}^p$ and ρ represents the correlation between predictors. We present some results for $p = 3$ and $n = 100$ in section 4.3.1. Following [31], we set the true data generating non-zero co-efficients to $\theta_i^* = \ln n$. We start by assuming one non-zero leading co-efficient ($\theta_1^* = \ln n$), then two and finally set all three co-efficients to be non-zero. We also vary this ordering of the non-zero co-efficients and present some results in appendix C Figures C.4-C.7. We allow ρ to vary between $\{0.0,\ 0.6,\ 0.8\}$.

4.3.1 Results

As before, we compare FAVI (FAVI_ssreg - algorithm 4.1), MF-VI and Gibbs sampling. Convergence for MF-VI and FAVI is determined by stabilization of the loss curves. For Gibbs sampling we use trace and auto-correlation plots. The kernel density plots we used in chapter 2 are useful representations for continuous probability distributions on $\mathbb{R}$, but are not effective visualizations for discrete distributions. Since in our case each θ_i is a mixture of a continuous

distribution and a discrete distribution at 0, we display the empirical cdf (ecdf) plots for the distribution of $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$. The empirical cdf at a point x for a random variable X on $\mathbb{R}$ is calculated as: $\hat{F}(x) = \sum_{i=1}^{N} \frac{\mathbb{I}\{X \leq x\}}{N}$. We use $N = 10,000$ samples to generate these plots. It is difficult to visualize central tendencies, variance and other important properties of the distribution on ecdf plots. Thus, for the case with a single non-zero leading co-efficient we include the kernel density plots, but remove the discrete samples at 0 for θ_i. We overlay bar plots onto the density plots to represent the proportion of zero samples. We also include violin plots in the Appendix C to aid the visualization of these distributional properties.[4]



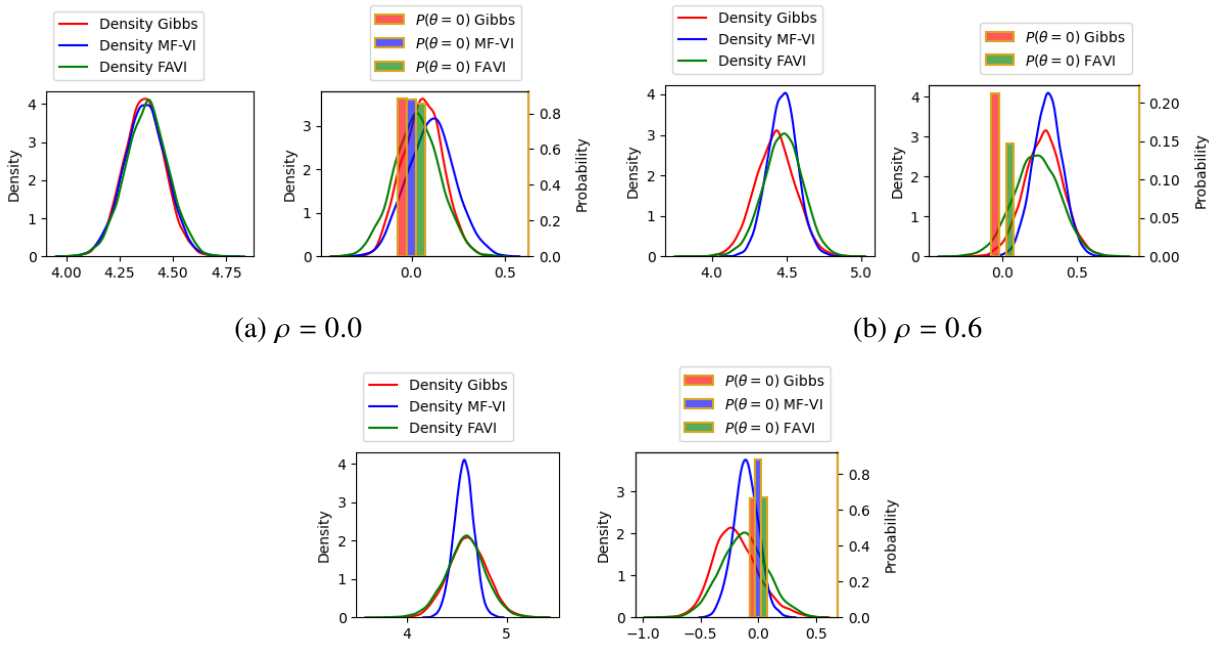
(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

Figure 4.1 Empirical cdf plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, 0, 0)$ and $n = 100$, $p = 3$.

Note that Gibbs sampling can be treated as the gold standard for recovery of the true posterior samples in this low-dimensional set-up. As expected, we see across Figures 4.1, 4.3 and 4.4 that an increase in ρ causes the distribution of MF-VI to deviate from that of Gibbs sampling, while FAVI closely mimics Gibbs sampling. There is one exception in Fig 4.1 ($\rho = 0.8$) where there is a small gap between FAVI and Gibbs sampling. We also see from the kernel density plots in Figure 4.2 that MF-VI underestimates the posterior variance as ρ increases. Further, in the kernel density plot 4.2 for the case $\rho = 0.6$ we see that MF-VI assigns a posterior inclusion probability (PIP) $\mathbb{P}(z_3 = 0) = 0$ as opposed to FAVI and MCMC which assigns a PIP $\mathbb{P}(z_3 = 0)$ between 0.15 and 0.2. These results suggest that FAVI is able to improve upon MF-VI in recovering the shape of the target posterior, particularly with respect to the posterior variance.

To further validate our conclusions we also present results for $n = 100$ with $p = 5, 10$. We set 20% of the leading co-efficients to $\ln n$, i.e. $\theta_i^* = \ln n$ for $1 \leq i \leq \lceil 0.2p \rceil$ and the remaining

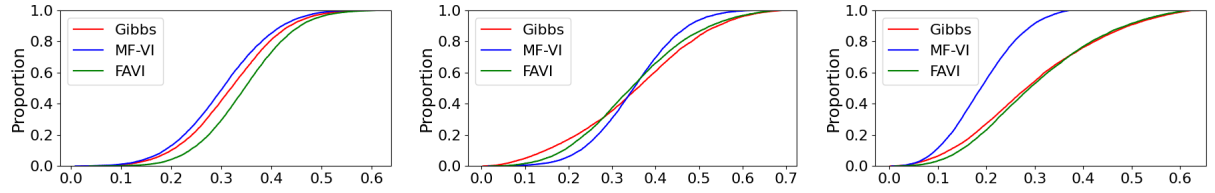
[4]See Table 3.1 for a recap of VI related notation used in the figure captions.



(a) $\rho = 0.0$ (b) $\rho = 0.6$

(c) $\rho = 0.8$

Figure 4.2 Kernel density plots for θ_1 (left) and θ_3 (right) where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, 0, 0)$ and $n = 100$, $p = 3$. Axis on the left is for the kernel density plots and axis on the right is for the bar plot.

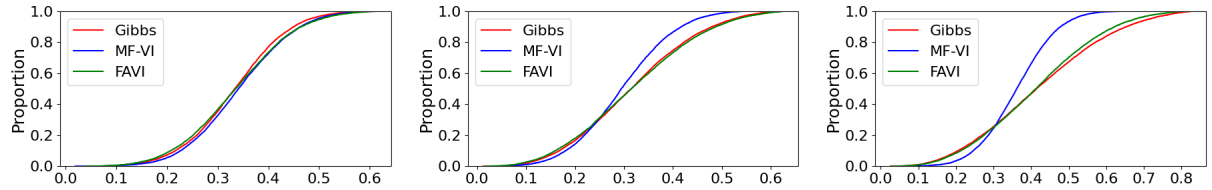


(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

Figure 4.3 Empirical cdf plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, \ln n, 0)$ and $n = 100$, $p = 3$.

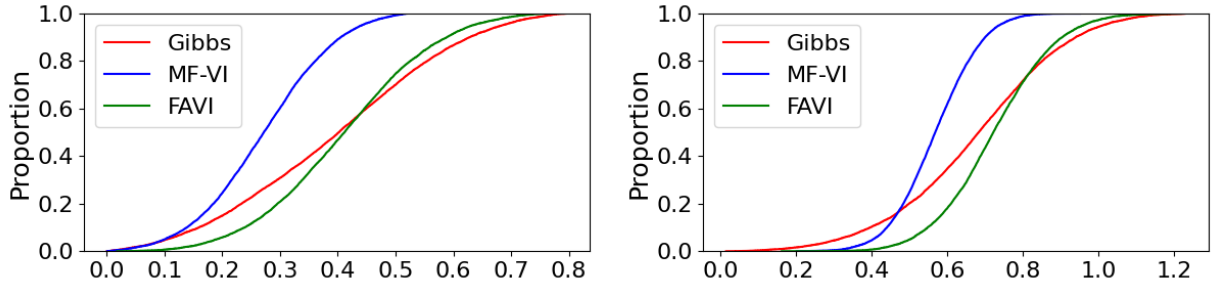
co-efficients to zero. We have already seen that for $\rho = 0$ there is no substantial difference for the 3 sampling methods. Additionally, as p increases having $\rho \geq 0.8$ is an extreme case in which PIP estimates even for Gibbs sampling may be inaccurate. Therefore, we use a moderate correlation of $\rho = 0.6$ for these cases. Results are presented in Figure 4.5.

We see that the results for $p = 5$ and $p = 10$ are largely consistent with that of $p = 3$. We make the additional observation that for $p = 10$ the discrepancy between FAVI and Gibbs increases slightly. This is confirmed by the violin plot presented in Figure 4.6. Thus, as the dimension p



(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

Figure 4.4 Empirical cdf plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, \ln n, \ln n)$ and $n = 100, p = 3$.



(a) $p = 5$ (b) $p = 10$

Figure 4.5 Empirical cdf plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $n = 100$. We set $\theta_i^* = \ln n$ for $1 \leq i \leq 0.2p$ and remaining $\theta_i^* = 0$. Here, $\rho = 0.6$.

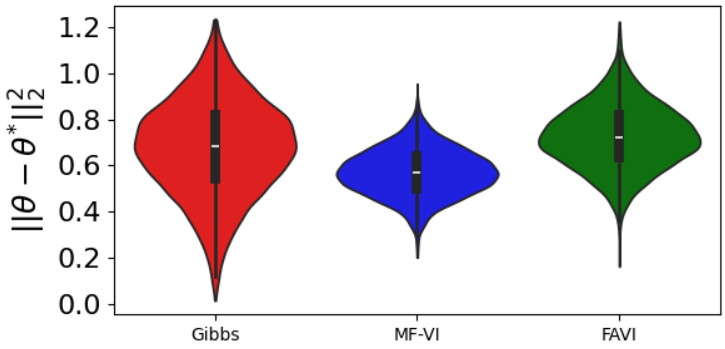


Figure 4.6 Violin plot for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $n = 100$ and $p = 10$. We set $\theta_i^* = \ln n$ for $1 \leq i \leq 0.2p$ and remaining $\theta_i^* = 0$. Here, $\rho = 0.6$.

increases, the ability of FAVI to recover the true posterior shape is limited by our assumption of a mean-field family for the discrete latent variables $\boldsymbol{z}$. We comment more on this in section 4.4. Run-time and accuracy comparisons between the 3 methods as p varies are presented in Table C.1 and Table C.2 in the appendix. Overall, the results are consistent with our earlier findings that FAVI scales more efficiently than MCMC and retains more attributes of the true posterior than MF-VI.

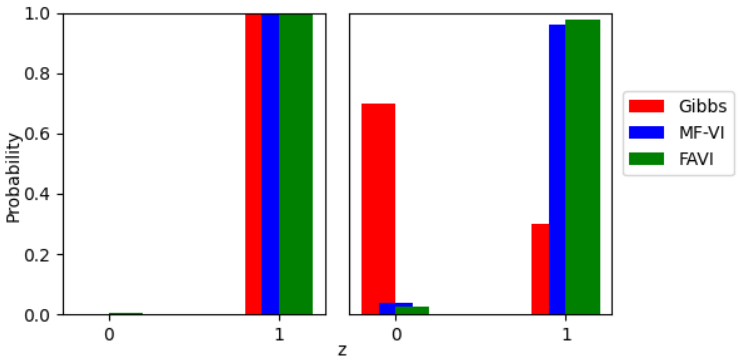


Figure 4.7 Bar plots for z_1(left), and z_3(right) where $\boldsymbol{z} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, IAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, \ln n, 0)$ and $n = 100, p = 3, \rho = 0.8$.

4.4 Limitations and future work

So far we have evaluated the approximate posterior from FAVI for $\boldsymbol{\theta}$, but what about the discrete latent variables $\boldsymbol{z}$? In Figure 4.7 we plot the distribution for z_1 and z_3 corresponding to the case when $\boldsymbol{\theta}_1^* = (\ln n, \ln n, 0)$ and $\rho = 0.8$. We see that both FAVI and MF-VI under-estimate the posterior variance for the marginal distribution of z_3 when compared to to Gibbs sampling. It is evident that while our algorithm can capture some of the dependencies across θ_i, it does not effectively model the dependency structure for $\boldsymbol{z}$. This is most likely due to our assumption of a mean-field family for $\boldsymbol{z}$.

A natural extension to the current algorithm would be to incorporate dependencies across z_i into our variational family. This problem is challenging because in order to have an exact model for the distribution of z_i, we require $O(2^p)$ variational parameters. One option is to use discrete flows for $\boldsymbol{z}$, but as discussed in section 2.7 the research in this area is nascent and current methods can only achieve a permutation of some base distribution. A more promising direction could be to use a continuous relaxation for the z_i's, such as the Gumble-softmax distribution and integrate them into the normalizing flow. We could also leverage our knowledge of the underlying statistical model dependency structure and relate those to the auto-regressive structure of the flows. For example with gaussianity assumptions for the slab, we know the complete conditional distributions. It can be verified that given z_i, θ_i only depends on $\theta_{-(i)}$ and similarly given θ_i, z_i only depends on $\theta_{-(i)}$. This additional information can be integrated into the conditioner networks to improve computational

efficiency.

Once these extensions are ironed out, we can proceed to experiments in higher dimensions and generalized linear models (glms). While our algorithm was outlined for a Gaussian likelihood, it can be easily modified to other likelihood functions since we use MC sampling (see 4.3). In fact, our experiments with logistic regression in chapter 2 showed that even for the simple case of $p = 2$, MF-VI was not able to recover the elliptical shape of the true posterior (see Figure 2.8a). This indicates that FAVI may demonstrate an improved gains over MF-VI for variable selection in glms. However, a much more rigorous empirical and theoretical study is required.

CHAPTER 5

CONCLUSIONS

With the onset of the big data era and increasing computational resources, developing scalable and accurate computational methods for Bayesian inference is more important than ever. This line of research derives much of its importance from its usefulness for uncertainty quantification, a critical component of answering many scientific questions.

In this dissertation, we have explored the use of Normalizing Flows aided Variational Inference (FAVI), a novel Bayesian computational tool that surfaced in contemporary machine learning literature. While it has been explored in many machine learning applications, this dissertation has focussed on analyzing FAVI from a statistical lens. We have focussed on fundamental theoretical questions relating to: (i) the expressivity of variational families generated by auto-regressive flows (IAF) (ii) the statistical accuracy vs scalability trade-off that is a feature of approximate inference methods and (iii) measures of uncertainty quantification.

Due to the challenging nature of these theoretical problems and the dearth of literature in the area, we have taken a bottom up approach to answer the questions outlined above. In chapter 2 we explored the use of FAVI for common statistical applications such as linear and logistic regression and provided comparisons to MCMC, which is very popular among statisticians and probabilists. Our experiments showed promise in FAVI as a tunable method that scales better than classical MCMC and provides improved accuracy in recovering the true posterior when compared to MF-VI. In chapter 3 we studied the approximate posterior resulting from using IAF within the context of the widely applicable Bayesian linear regression model. We have derived theoretical results on the behaviour of the optimal Kullback-Leibler divergence and the loss in uncertainty quantification from the variational approximation. We provided intuitive explanations for these results relating them to correlation between the regression predictors and eigen values of the covariance matrix. To our knowledge we are the *first* to provide such a theoretical analysis and the results are novel in their approach to analyzing uncertainty quantification for variational inference. We believe this can open up many avenues of future research. Finally, we have proposed an algorithm leveraging FAVI

for Bayesian variable selection with spike and slab regression. Preliminary experiments show that FAVI has the potential to effectively capture dependencies in the posterior distribution.

Considerable future work is required to make FAVI more widely accepted for statistical inference. Expanding our proofs, we hope to explore polynomial activations in the conditioner networks and leverage their universal approximation properties to tackle non-linear activations. Further, we could go beyond the linear model to generalized linear models (glms), where the form of the true posterior is unknown and other non-Gaussian posteriors. We can also consider extensions from IAF to Neural Auto-regressive Flows [16] (which are universal approximators) for our theoretical results. For the Bayesian spike and slab regression method we proposed in chapter 4, more expansive, large scale experimentation can be conducted as a first step. Following this we can consider extensions to glms and incoporate dependencies into the discrete latent variables.

We expect to see future collaborations between computer scientists and statisticians to address some of these interesting and impactful open problems.

BIBLIOGRAPHY

[1] Maxime Beauchamp. On numerical computation for the distribution of the convolution of n independent rectified gaussian variables. *Journal de la Société de statistique de Paris*, 159:88–111, 03 2018.

[2] Kush Bhatia, Nikki Kuang Lijing, Yi-An Ma, and Yixin Wang. Statistical and computational trade-offs in variational inference: A case study in inferential model selection, 2023.

[3] Shrijita Bhattacharya and Tapabrata Maiti. Statistical foundation of variational bayes neural networks. *Neural Networks*, 137:151–173, 2021.

[4] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.

[6] Léon Bottou. Online algorithms and stochastic approximations. In David Saad, editor, *Online Learning and Neural Networks*. Cambridge University Press, Cambridge, UK, 1998. revised, oct 2012.

[7] Peter Carbonetto and Matthew Stephens. Scalable Variational Inference for Bayesian Variable Selection in Regression, and Its Accuracy in Genetic Association Studies. *Bayesian Analysis*, 7(1):73–108, 2012.

[8] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

[9] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335, 1995.

[10] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

[11] James E. Gentle. *Numerical Linear Algebra*, pages 203–240. Springer New York, New York, NY, 2009.

[12] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. Made: Masked autoencoder for distribution estimation. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 881–889, Lille, France, 2015.

[13] Yongtao Guan and Matthew Stephens. Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815, 2011.

[14] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, 15(47):1593–1623, 2014.

[15] Matthew D. Hoffman, Pavel Sountsov, Joshua V. Dillon, Ian Langmore, Dustin Tran, and Srinivas Vasudevan. Neutra-lizing bad geometry in hamiltonian monte carlo using neural transport. *arXiv: Computation*, abs/1903.03704, 2019.

[16] Chin-Wei Huang, David Krueger, Alexandre Lacoste, and Aaron Courville. Neural autoregressive flows. In *Proceedings of the 35th International Conference on Machine Learning (ICML-2018)*, 2018.

[17] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017.

[18] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

[20] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[21] Diederik P. Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. In *Proceedings of the 30th Advances in Neural Information Processing Systems (NeurIPS 2016)*, page 4743–4751, Barcelona, Spain., 2016.

[22] Ivan Kobyzev, Simon Prince, and Marcus Brubaker. Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1, 05 2020.

[23] Michael Komodromos, Marina Evangelou, Sarah Filippi, and Kolyan Ray. Group spike and slab variational bayes, 2023.

[24] Yingzhen Li and Richard E. Turner. Rényi divergence variational inference. In *NIPS*, 2016.

[25] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *5th International Conference on Learning Representations (ICLR 2017)*, Toulon, France, 2017.

[26] Oren Mangoubi, Natesh S. Pillai, and Aaron Smith. Does hamiltonian monte carlo mix faster than a random walk on multimodal densities? *arXiv*, 2018.

[27] ROBERT W. NEWCOMB. On the simultaneous diagonalization of two semi-definite matrices. *Quarterly of Applied Mathematics*, 19(2):144–146, 1961.

[28] George Papamakarios, Eric Nalisnick, Danilo J. Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[29] Sumegha Premchandar, Shrijita Bhattacharya, and Tapabrata Maiti. Normalizing flows aided variational inference: A useful alternative to mcmc? *Notices of the American Mathematical Society*, 70(7):1059–1070, 2023.

[30] Neal M. Radford. *MCMC using Hamiltonian dynamics*, chapter 5. Chapman and Hall (CRC Press), 2011.

[31] Kolyan Ray and Botond Szabo. Variational bayes for high-dimensional linear regression with sparse priors. *Journal of the American Statistical Association*, 117:1–31, 11 2020.

[32] Kolyan Ray, Botond Szabo, and Gabriel Clara. Spike and slab variational bayes for high dimensional logistic regression. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14423–14434. Curran Associates, Inc., 2020.

[33] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

[34] Herbert Robbins and Sutton Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951.

[35] Vivekananda Roy. Convergence diagnostics for markov chain monte carlo. *Annual Review of Statistics and Its Application*, 2019.

[36] Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Science/Engineering/Math, 1986.

[37] Axel Ruhe. Perturbation bounds for means of eigenvalues and invariant subspaces. *BIT Numerical Mathematics*, 10:343–354, 1970.

[38] Esteban G. Tabak and Cristina Vilma Turner. A family of nonparametric density estimation algorithms. *Communications on Pure and Applied Mathematics*, 66, 2013.

[39] Esteban G. Tabak and Eric Vanden-Eijnden. Density estimation by dual ascent of the log-likelihood. *Communications in Mathematical Sciences*, 8:217–233, 2010.

[40] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[41] Gao Wang, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(3):1273–1300, 2020.

[42] Yixin Wang and David M. Blei. Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161, 2019.

[43] Yu Wang, Fang Liu, and Daniele E. Schiavazzi. Variational inference with nofas: Normalizing flow with adaptive surrogate for computationally expensive models. *Journal of Computational Physics*, 467:111454, 2022.

[44] Stefan Webb, Jonathan P. Chen, Martin Jankowiak, and Noah Goodman. Improving automated variational inference with normalizing flows. In *ICML Workshop on Automated Machine Learning*, 2019.

[45] Cheng Zhang. Improved variational bayesian phylogenetic inference with normalizing flows. In *Advances in Neural Information Processing Systems*, volume 33, pages 18760–18771. Curran Associates, Inc., 2020.

[46] Xuebin Zhao, Andrew Curtis, and Xin Zhang. Bayesian seismic tomography using normalizing flows. *Geophysical Journal International*, 228:213–239, 2021.

APPENDIX A

INFORMATION ON ENERGY DENSITY FUNCTIONS

In this section we define energy density functions $U3$, $U4$ and $U9$ used in our experiments comparing FAVI and the RW-MH algorithms in Section 2.6.2 of this dissertation. We specifically define these functions due to the pattern displayed in our experiments wherein FAVI substantially outperforms RW-MH in run-time. The remaining energy density functions can be found in toy_energy.

Let $\boldsymbol{z} = (z_2, z_2)$, then the density for $U3$ is given by:

$$U3(\boldsymbol{z}) = \exp\left\{ -\left(2 - \sqrt{z_1^2 + 0.5z_2^2}\right)^2 \right\} \tag{A.1}$$

The density $U4$ is a mixture of 4 Gaussians with means $\mu_1 = (-5, 0)$, $\mu_2 = (5, 0)$, $\mu_3 = (0, 5)$, $\mu_4 = (0, -5)$ and variance $\sigma^2 = 1.5$.

$$U4(\boldsymbol{z}) = 0.1\frac{e^{-\frac{1}{2\sigma^2}(\boldsymbol{z}-\mu_1)^\top(\boldsymbol{z}-\mu_1)}}{\sqrt{2\pi}\sigma} + 0.3\frac{e^{-\frac{1}{2\sigma^2}(\boldsymbol{z}-\mu_2)^\top(\boldsymbol{z}-\mu_2)}}{\sqrt{2\pi}\sigma} + 0.4\frac{e^{-\frac{1}{2\sigma^2}(\boldsymbol{z}-\mu_3)^\top(\boldsymbol{z}-\mu_3)}}{\sqrt{2\pi}\sigma}$$
$$+ 0.2\frac{e^{-\frac{1}{2\sigma^2}(\boldsymbol{z}-\mu_4)^\top(\boldsymbol{z}-\mu_4)}}{\sqrt{2\pi}\sigma}$$

The density $U9$ is a complex mixture of multiple densities and is provided below.

$$U9(\boldsymbol{z}) = f_1(\boldsymbol{z}) + f_2(\boldsymbol{z})$$
$$f_1(\boldsymbol{z}) = exp\left\{ \ln\left(e^{-0.5((z_2-w_1)/0.4)^2} + e^{-0.5((z_2-w_1+w_3)/0.35)^2}\right) - 0.05(z_1^2 + z_2^2)\right\}$$
$$w_1 = \sin(\frac{1}{2}z_1\pi)$$
$$w_3 = 2.5 \times 1/(1 + e^{-((z_1-2)/0.3)})$$
$$f_2(\boldsymbol{z}) = \exp\{2 * U3(\boldsymbol{z} * 1.5 - 2)\}$$

APPENDIX B

RUN-TIME COMPARISONS FOR LOGISTIC REGRESSION

Avg. Time to converge $\pm$ s.d (seconds)			
n, p	**MH**	**FAVI**	**MF-VI**
$(50, 2)$	43 ± 1	169 ± 4	155 ± 2
$(50, 20)$	206 ± 5	265 ± 21	158 ± 11
$(50, 50)$	**423** ± 11	**371** ± 16	157 ± 4
$(50, 100)$	**853** ± 44	**496** ± 106	151 ± 8
$(100, 2)$	44 ± 1	183 ± 19	161 ± 10
$(100, 20)$	227 ± 25	294 ± 37	166 ± 17
$(100, 50)$	**876** ± 20	**380** ± 16	160 ± 4
$(100, 100)$	**1167** ± 93	**472** ± 41	167 ± 17
$(200, 2)$	45 ± 8	183 ± 23	164 ± 25
$(200, 20)$	233 ± 30	276 ± 33	172 ± 17
$(200, 50)$	**921** ± 16	**375** ± 13	157 ± 7
$(200, 100)$	**1223** ± 76	**471** ± 23	172 ± 12

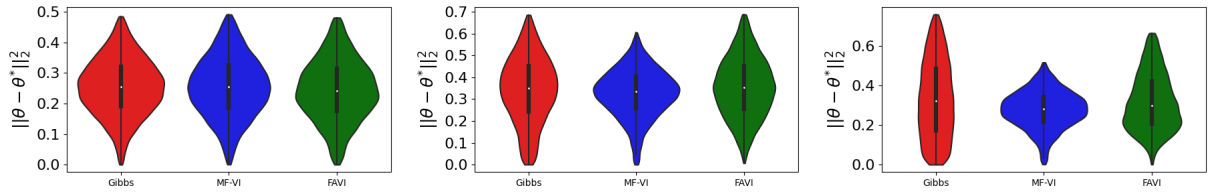
Table B.1 *Avg algorithm runtime ± s.d. for logistic regression experiments in Section 2.6.4. Results are computed over 5 trials (smaller values are better).*

APPENDIX C

ADDITIONAL RESULTS FOR SPIKE AND SLAB REGRESSION

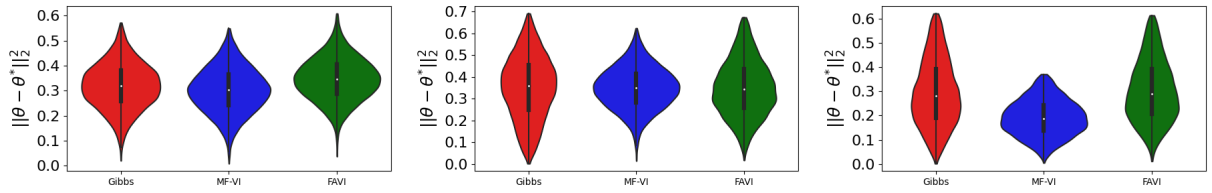
Remark: In order to effectively compare the 3 methods via violin plots we remove outliers (values greater than $1.5 \times 75\%$ quantile). This is because MCMC produces a few very extreme samples that makes the scale of the 3 distributions appear very different and it becomes difficult to visualize all of them simultaneously if those samples are not excluded.

The violin plots presented in Figures C.1, C.2 and C.3 below are for the case $n = 100, p = 3$ where we set leading coefficients non-zero, from one non-zero co-efficient to all three non-zero.



(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

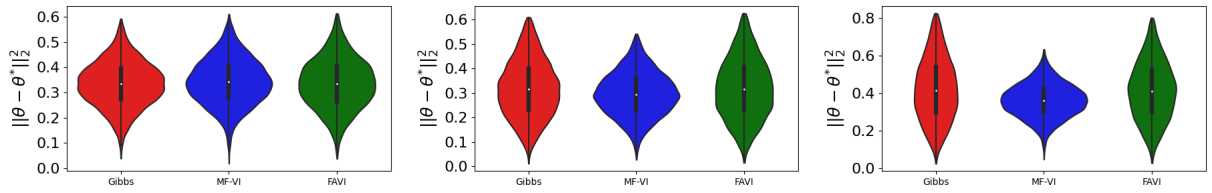
Figure C.1 violin plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, 0, 0)$ and $n = 100, p = 3$.



(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

Figure C.2 Violin plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ ($\cdot$ is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, \ln n, 0)$ and $n = 100, p = 3$.

We also present results for $n = 100, p = 3$ but change the ordering of the non-zero coefficients. We display all possible orderings but only for correlation $\rho = 0.6$. Ecdf and violin plots are presented in Figures C.4 - C.7.



(a) $\rho = 0.0$ (b) $\rho = 0.6$ (c) $\rho = 0.8$

Figure C.3 Violin plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, \ln n, \ln n)$ and $n = 100$, $p = 3$.

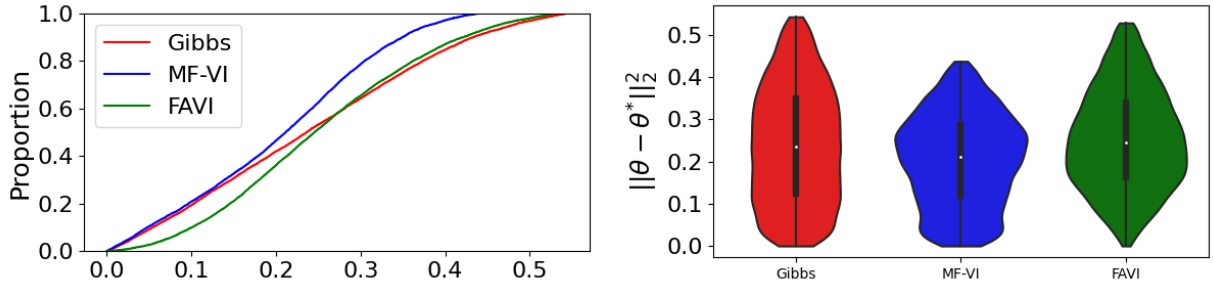


Figure C.4 Ecdf (left) and violin plots (right) for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (0, \ln n, 0)$ and $n = 100$, $p = 3$.

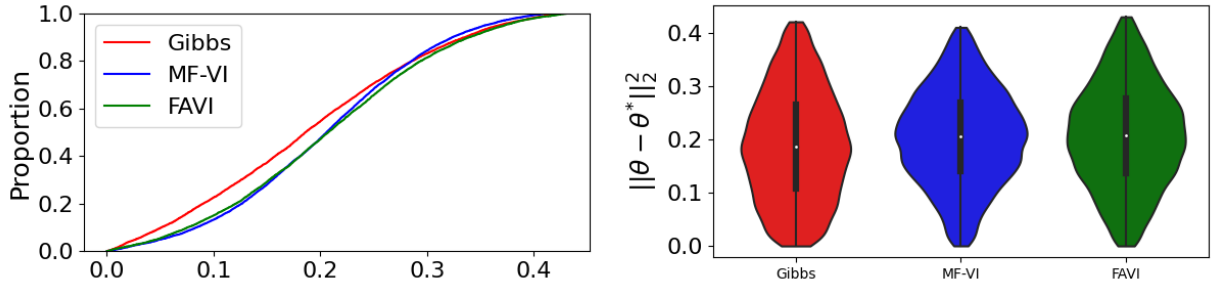


Figure C.5 Ecdf (left) and violin plots (right) for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (0, 0, \ln n)$, $\theta_i^* = 0$ for $i = 1, 2$ and $n = 100$, $p = 3$.

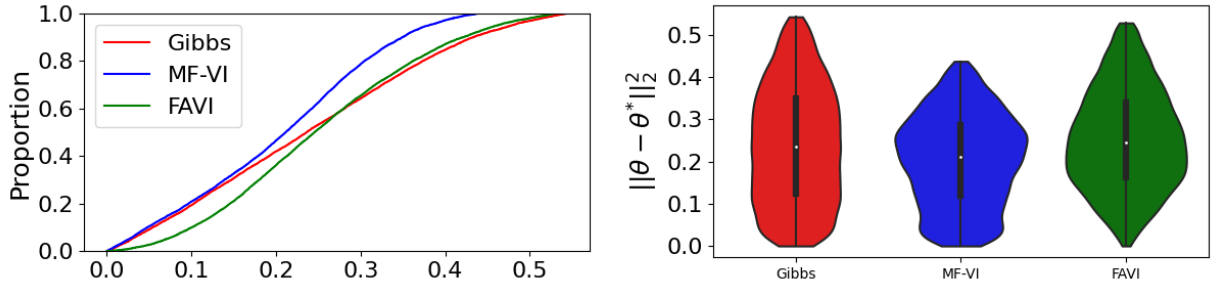


Figure C.6 Ecdf (left) and violin plots (right) for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{.\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (\ln n, 0, \ln n)$ and $n = 100$, $p = 3$.

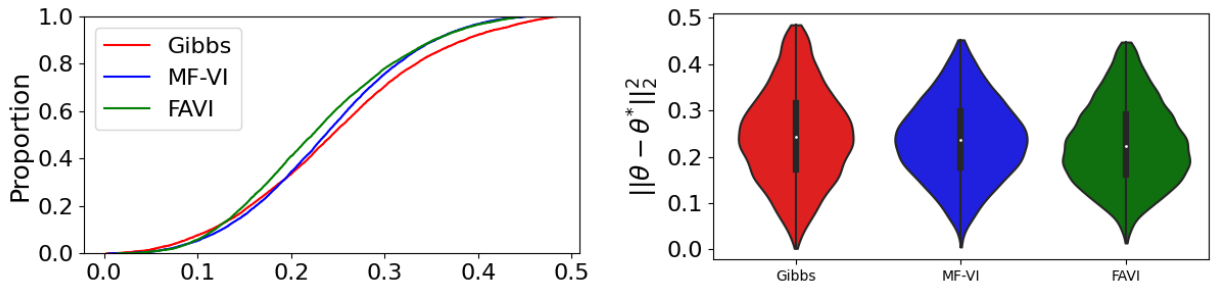


Figure C.7 Ecdf (left) and violin plots (right) for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ (. is MF, NAF, Gibbs) when $\boldsymbol{\theta}^* = (0, \ln n, \ln n)$ and $n = 100$, $p = 3$.

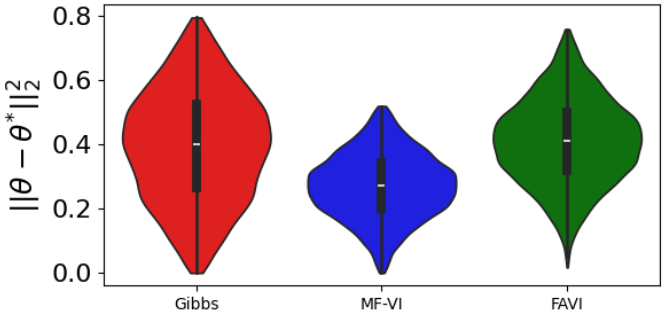


Figure C.8 Empirical cdf plots for $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2^2$ where $\boldsymbol{\theta} \sim q_{\phi^*}^{\{\cdot\}}$ (. is MF, NAF, Gibbs) when $n = 100$. We set $\theta_i^* = \ln n$ for $1 \leq i \leq 0.2p$ and remaining $\theta_i^* = 0$. We have $p = 5$ and $\rho = 0.6$.

Avg. algorithm run-time in seconds.			
p	**Gibbs**	**MF-VI**	**FAVI**
5	4.6	3.8	3.0
10	27	3.9	2.9
15	30.9	3.8	3.1
20	30.9	3.9	3.2
30	63.4	4	3.2
40	42.5	4.5	3.1
50	59.3	4.2	3.7
70	85.2	5.1	3.5
90	148.4	4.3	3.8

Table C.1 *Avg algorithm runtime for spike and slab regression experiments in Section 4.3.1 (smaller values are better). We set* $\theta_i^* = \ln n$ *where* $n = 100$ *for* $1 \leq i \leq \lceil 0.2p \rceil$. *We choose 16 nodes in* c_ϕ *and 4 nodes in the DSF. Here,* $\rho = 0.6$.

	AUC PR		
p	**Gibbs**	**FAVI**	**MF-VI**
5	0.5	0.5	0.5
10	0.67	0.67	0.67
15	0.75	0.75	0.75
20	0.8	0.8	0.8
30	0.86	0.86	0.86
40	0.89	0.89	0.89
50	0.91	0.91	0.91
70	0.93	0.93	0.93
90	0.95	0.95	0.95

Table C.2 *AUC Precision-Recall (larger values are better) for spike and slab regression experiments in Section 4.3.1. We set* $\theta_i^* = \ln n$ *where* $n = 100$ *for* $1 \leq i \leq \lceil 0.2p \rceil$. *We choose 16 nodes in* c_ϕ *and 4 nodes in the DSF. Here,* $\rho = 0.6$.