Bayesian Inference for Non-linear forward model by using a VAE-based neural network structure

Yechuan Zhang[1], Jian-Qing Zheng[2], and Michael Chappell[*]

[1]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford
[2]The Kennedy Institute of Rheumatology, Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences, University of Oxford
[*]Sir Peter Mansfield Imaging Centre, School of Medicine, University of Nottingham
[*]Mental Health & Clinical Neurosciences, School of Medicine, University of Nottingham
[*]Neuroimaging, FMRIB, Nuffield Department of Clinical Neurosciences, University of Oxford

***Abstract*—In this paper, a Variational Autoencoder (VAE) based framework is introduced to solve parameter estimation problems for non-linear forward models. In particular, we focus on applications in the field of medical imaging where many thousands of model-based inference analyses might be required to populate a single parametric map. We adopt the concept from Variational Bayes (VB) of using an approximate representation of the posterior, and the concept from the VAE of using the latent space representation to encode the parameters of a forward model. Our work develops the idea of mapping between time-series data and latent parameters using a neural network in variational way. A loss function that differs from the classic VAE formulation and a new sampling strategy are proposed to enable uncertainty estimation as part of the forward model inference. The VAE-based structure is evaluated using simulation experiments on a simple example and two perfusion MRI forward models. Compared with analytical VB (aVB) and Markov Chain Monte Carlo (MCMC), our VAE-based model achieves comparable accuracy, and hundredfold improvement in computational time (100ms/image). We believe this VAE-like framework can be generalized to imaging modularities with higher complexity and thus benefit clinical adoption where otherwise long processing time associated with conventional inference methods is prohibitive.**

***Keywords*—** Variational Autoencoder, Variational Bayes, Parameter Estimation Problems, Medical Imaging

I. Introduction

BAYESIAN inference is a popular statistical technique which can be used to estimate parameters of a generative model from data. It has been used in a wide variety of applications, from pricing decision making in the field of business and commerce [1] to disease mapping and medical diagnostics [2]. In particular, it has been applied to parameter mapping problems in MRI, including segmentation of sub-cortical structure [3], inference on functional MRI (FMRI) time-series data [5] and optimization of the haemodynamic response function (HRF) in FMRI [4]. Unlike frequentist statistical approaches, Bayesian inference introduces the concept of a posterior probability distribution, enabling parameter estimation along with associated measures of confidence. Bayesian inference involves belief updating and allows for the specification of prior probability distributions. This is particularly helpful in parameter mapping problems with noisy data where prior information about the parameters is known. For example, Bayesian inference allows the incorporation of physiologically plausible ranges of biophysical parameters as prior information into the perfusion parameter estimation problem in Arterial Spin Labelling (ASL) MRI [6] where ASL is well established as a quantitative technique to measure perfusion. In principle, Bayes's theory enables the computation of a posterior distribution; however, because of the intractable integral that typically arises, Bayesian methods often do not provide results in closed form.

Multiple approaches have been proposed to approximate Bayesian computation. For example, the Laplace approximation [10] [11] assumes a posterior distribution with a Gaussian distribution centered at the maximum a posteriori (MAP) parameter estimate and Markov Chain Monte Carlo (MCMC) [12] [13] achieves Bayesian estimates by sampling from a probability distribution and creating a Markov Chain. However, the form of the true posterior may not exhibit like a Gaussian likelihood and MCMC usually requires an extremely high computational cost. Both of the limitations can be problematic in high dimensional applications like computation across many voxels in medical imaging data, precluding deployment in clinical applications where rapid computation is called for. To help with these scenarios, fast computation techniques are required. Variational Bayes (VB) [8] [9] attempts to achieve a balance between these two factors. VB methods optimize the posterior distribution estimation by minimizing Kullback Leibler (KL) divergence [14] between true posterior and an approximate posterior. The original VB implementation, analytical VB (aVB) [9] requires a tractable integral in the KL divergence computation which restricts the prior and posterior to be from the conjugate exponential family. Conventional aVB methods only work for linear forward models, but recent research papers generalized them to non-linear models [6]. More recent variants, such as Stochastic VB (sVB) [7] eases the limitation in the form of posterior distribution via introduction of stochastic gradient descent to directly perform optimization.

The optimization process in both aVB and sVB is iterative,

resulting in an estimator that scales in computational cost with the complexity of the forward model. Moreover, the optimization process is "memoryless" – every time a new data set is received, a completely new optimization needs to be conducted to find the optimal measurements of model parameters. The development of neural networks provides a possible solution to this problem by amortizing the repetitive computational cost produced by new data sets among the neural network training process. For example, Lahiri et al. [15] used a regression neural network to replace the conventional ASL magnetic resonance fingerprinting (MRF) [16] dictionary mapping, reducing the computational time of estimating a single-slice data to few seconds. Luciw et al. [17] trained a convolutional neural network (CNN) [18] and a U-Net [19] with real 3D data sets for automated perfusion estimation. However, these applications only provided conventional parameter estimations without uncertainty measures. Moreover, using real data as training data limits generalization to other data sets in a way not seen for existing VB algorithms that infer directly using the forward model.

Variational Bayes and neural network representations have been previously combined into the Variational Autoencoder (VAE) [20]. The conventional VAE is widely accepted as a dimension reduction and image denoising technique [21]. Bliesener et al. [22] in 2020 proposed a neural network combined with a variational loss to perform parameter estimation on time-series DCE-MRI data, their approach partly mirrors that of a VAE but only included an encoder and did not consider an encoder-decoder pair. In this work, we combine encoder and decoder and propose a VAE-like neural network structure to perform Bayesian inference for the parameter of a non-linear forward model. Like Bliesener [22], we directly linked the latent parameters to models parameters. Crucially, the decoder is specified as the forward model. The VAE-based neural network was compared with MCMC and aVB on both a representative 'toy' example and practical examples in ASL perfusion MRI [23] [24], showing consistency among the different methods and offering orders of magnitude improvements in calculation for inference on new data using neural networks. The main contributions of our work are:

- A Bayesian inference strategy based on a neural network architecture that can be applied for parameter and uncertainty estimation for any non-linear forward model from serial data.
- Comparable accuracy in parameter inference and a hundred-fold improvement in computational speed compared to MCMC and a conventional VB approach.
- A new loss function and a novel sampling strategies that simplify the calculation of loss and enable network training.

In the rest of the paper, we will discuss the theory behind our work (Section II), methods we proposed (Section III), experiment design and results (Section IV) and finally the achievements and limitations (Section V and VI).

II. Theory

A. Bayesian Inference

Bayesian inference is an approach to determine the posterior distribution for the parameters of a generative (forward) model $p(\Theta|\mathbf{Y})$ using measured data. According to Bayes theorem, the posterior distribution can be written as:

$$p(\Theta \mid \mathbf{Y}, \mathcal{M}) = \frac{p(\mathbf{Y} \mid \Theta, \mathcal{M}) p(\Theta \mid \mathcal{M})}{p(\mathbf{Y} \mid \mathcal{M})} \tag{1}$$

where $\mathcal{M}$, $\mathbf{Y}$ and Θ represent model, data and parameters of interest respectively. $P(\Theta)$ and $P(\mathbf{Y}|\theta)$ indicate prior probability and data likelihood. The expected value and dispersion measures of $p(\Theta|\mathbf{Y})$, e.g., the mean and variance, provide information on the best estimates of θ and associated uncertainties.

B. Bayesian Inference for Non-linear forward models

The relationship between parameters θ and data y can be formulated by a non-linear generative model g.

$$y = g(\theta) + \epsilon \tag{2}$$

where $\boldsymbol{y}$ is a series data vector of length K, g is a non-linear generating function, θ is a vector of parameters of interest and ϵ indicates an error term. Under the assumption that the data is corrupted by Gaussian white noise, this term follows a Gaussian distribution i.e. $\epsilon \sim \mathcal{N}(\mathbf{0}, e^2\mathbf{I})$ where e is the noise parameter, the standard deviation of the error term ϵ. The resulting data likelihood of y is a Gaussian distribution, where the log-likelihood can be written as:

$$\log P(y \mid \theta) = -K \log e - \frac{1}{2} e^{-2} (y - g(\theta))^{\mathrm{T}} (y - g(\theta)) \tag{3}$$

where $\Theta = \{\theta, e\}$ is the set of all parameters of interest.

C. Variational Bayesian Inference

Bayes theorem allows the calculation of posterior distribution, however, due to the integral involved in the calculation of $p(\mathbf{Y} \mid \mathcal{M})$ in equation (1), in most cases, analytical solutions for posterior distribution cannot be derived.

Variational Bayes (VB) has been proposed to allow evaluation for an analytical solutions for the posterior by introducing an approximate posterior $q(\theta)$ [8]. The difference between $q(\theta)$ and the true posterior distribution $P(\theta|y)$ can be quantified by Kullback–Leibler (KL) divergence. The optimal approximate posterior is achieved by the minimization of KL divergence which is equivalent to the maximization of free energy F, also known as the Evidence Lower Bound (ELBO):

$$F = \int q(\mathbf{w}) \log \left[\frac{P(\mathbf{y} \mid \mathbf{w}) P(\mathbf{w})}{q(\mathbf{w})} \right] d\mathbf{w} \tag{4}$$

One common option in VB is the mean field approximation which divides model parameters into subgroups and assumes the independence of posterior distributions for each subgroup. The posterior distribution for all model parameters can thus be factorized to the multiplication of individual groups. The

additional assumption of conjugate prior allows the approximate posterior to be written in closed form. We refer to this approach as analytical VB. In another approach, stochastic VB, stochastic gradient descent is applied directly to optimisation of the free energy without any limitations in form of the distributions. These two approaches both result in an iterative update process for the hyper parameters of the approximate posterior distribution $q(\theta)$, whilst these schemes eventually converge to an optimal solution, both may still have a high computational cost if the evaluation of the forward model is not trivial.

D. Variational Autoencoder

The Variational Autoencoder is a deep learning architecture composed of three parts: an encoder, a decoder and a variational latent space between the two. It is often used for signal denoising problems, where the encoder and decoder are two neural networks used for dimension reduction and dimension expansion respectively. Unlike the conventional Autoencoder, the VAE latent space is not simply a vector of parameter values, but a predifine multivariate distribution. The variational layer imposes a constrain on latent distribution which acts as a form of regularisation. This architecture is trained to minimise the reconstruction error between the recovered (i.e., encoded-decoded) signal and input signal. The loss function is usually adopted as the KL-divergence between true posterior of θ $P = p(\theta|y)$ and the approximate posterior of θ , $i.e., Q = q(\theta|y)$:

$$\mathrm{KL}\left(Q||P\right) = \int q\left(z \mid y\right) \log \frac{q\left(z \mid y\right)}{p\left(z \mid y\right)} dz \tag{5}$$

By simple calculation, minimizing $\mathrm{KL}\left(Q||P\right)$ is equivalent to :

$$\min \mathbb{E}_q[\log q(\theta \mid y) - \log p(\theta)] - \mathbb{E}_q[\log p(y \mid \theta)] \tag{6}$$

The first term in equation (6) is a KL-loss between posterior $q(\theta|y)$ and prior $p(\theta)$, which enforces the approximate posterior to be similar to prior distribution. The second term is called reconstruction loss, enforcing data consistency between input signal and recovered signal. In a classic VAE application, θ is simply a lower-dimensional representation in latent space without the parameter(s) having any particular interpretation.

III. Methods

In this paper we adopt a VAE-based structure to train a network that can directly perform Bayesian inference on series data using a non-linear forward model without any repetition of the optimization process for each new dataset. The VAE-based method is adopted from a conventional VAE structure [20] (Section III-A), trained with a new loss function (Section III-B) and a novel learning strategy (Section III-C).

A. Structure of VAE-like Neural Network

Figure 1 illustrates how a VAE-like framework can be used to train a neural network (the encoder in figure 1) to perform Variational Bayesian inference. This formulation is generic to problems with a series data-model combination. Similar to a conventional VAE, the VAE-like structure contains three parts: an encoder network, a variational layer defining the latent distributions for parameter of interest and, taking the place of the decoder, a forward model. Unlike the conventional VAE, the latent variables of VAE-like structure are specifically identified with generative model parameters and the first-and-second order samplers implemented to generate samples from latent layer to calculate the loss. The loss function consists of two reconstruction loss and two distribution loss terms, discussed in the following section. In our neural network architecture, the input layer, the encoder and the latent layer combined form the final inference network that can be used to performed inference on new data, the rest of elements are used to train the network.

B. Loss Function in Training VAE-like Neural Network

As shown in equation 6, the conventional VAE loss consists of a reconstruction loss term which measures the consistency between inputs and outputs and a prior-posterior loss term which measures the similarity between prior and posterior distributions. The loss of VAE-like framework is designed based on the conventional VAE loss. This loss of VAE-like framework is the sum of a Kullback Leibler (KL) loss between prior and posterior (i), a reconstruction loss between input and output (ii), a KL loss between input and output (iii) and a regression loss term over parameters of interest (iv), where the first two terms added together are the conventional loss.

The uncertainty in estimation for a serial data-model combination comes from two sources: the intrinsic variability of parameters and the noise introduced by data acquisition. The first source of uncertainty can be captured by the first KL loss (equation (7)(i)) which was used to guarantee the similarity between posterior distribution $q(\theta \mid y)$ and prior belief. But the second is not included in the conventional loss expression. Thus, the second KL loss term was added for our implementation. The second KL loss (equation (7)(iii)) measured the discrepancy between distribution of input data $P(y)$ and distribution of recovered data $P'(y')$. The reconstruction term (ii) and regression loss term (iv) enforce the data consistency between input and output and the consistency between true parameters and network inferred parameters respectively. By adding the regression loss term (iv) and KL loss term (iii), the consistency loss and distribution loss were symmetric between data and parameter of interest. The mean square error (MSE) or relative absolute error (RAE) were considered as the regression loss term. The coefficient λ is given as a scaling factor to regression loss term.

$$\begin{aligned}\mathcal{L} &= \underbrace{\mathrm{KL}(q(\theta \mid y)||p(\theta))}_{\text{(i) prior}-\text{posterior}} - \underbrace{\mathbb{E}_q[\log p(y \mid \theta)]}_{\text{(ii) reconstruction}} + \underbrace{\mathrm{KL}(P(y)||P'(y'))}_{\text{(iii) recovered signal}} \\ &\quad + \underbrace{\lambda \mathrm{Reg}\left(\mu(\theta), \theta^{\mathrm{true}}\right)}_{\text{(iv) regression}} \\ &= \mathbb{E}_q[\log q(\theta \mid y) - \log p(\theta)] - \mathbb{E}_q[\log p(y \mid \theta)] \\ &\quad - \mathbb{E}_p\left[\log p(y) - \log p'(y')\right] + \lambda \mathrm{Reg}\left(\mu(\theta), \theta^{\mathrm{true}}\right)\end{aligned} \tag{7}$$

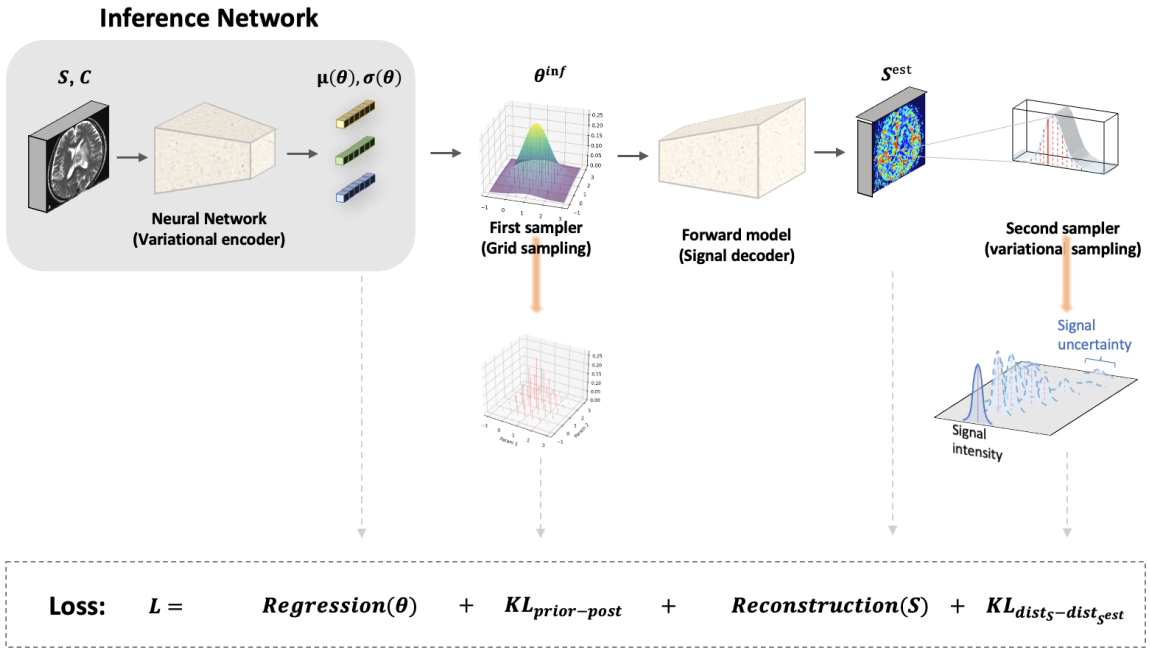

Fig. 1. Structure of VAE-like Framework: This figure shows the structure of our VAE-like framework. The inference network (grey shading) consists of an input layer, an encoder and a latent distribution. The input consists of time series data and known parameters with dimension $R^{n\times(m_1+m_2)}$, where m_1 and m_2 denotes the number of post-label delays and number of known parameters. The output are the hyperparameters of latent distributions, normally the mean μ and standard deviation σ of θ and noise parameter e, where $\mu, \sigma \in R^{n\times(|\theta|+|e|)}$, where $|.|$ denotes cardinality. The rest of the elements including samplers, forward model and estimated output are used in neural network training. The loss function consisting of a reconstruction loss, a regression loss and two KL loss used in training process is highlighted in the bottom of the figure. The two inputs of variational encoder, the time-series data and known parameters of model are denoted as $\boldsymbol{S}$ and $\boldsymbol{C}$. The hyperparameters of latent distribution are denoted as $\mu(\theta)$ and $\sigma(\theta)$ and parameter inference is denoted as $\theta^{\inf}$.

C. First-and-Second Order Sampler

As shown in equation (7), the computation of the first three loss terms requires integration over continuous distribution of θ. However, because of the intractability of these integrals, integration approximation methods are needed. A first-and-second-order sampler was proposed to address this issue in an efficient way. The first-order sampler was designed to discretize the sample space of parameter θ and the noise parameter e while the second-order sampler was used to generate realizations of noise.

The calculation of the reconstruction term and the KL loss between prior and posterior in loss function (7)(i) and (7)(ii) only required the first-order sampling. Equation (7)(i) can be approximated by:

$$\begin{aligned}&\mathbb{E}_q[\log q(\theta \mid y) - \log p(\theta)]\\ &= \int q(z)[\log q(z \mid y) - \log p(z)]dz\\ &\approx \sum_{\Theta^*\in\Phi,\Theta^*=(\theta^*,e^*)} q^*(\theta^*)[\log q(\theta^*|y) - \log p(\theta^*)]\end{aligned} \tag{8}$$

By assuming the data likelihood to be a Gaussian distribution, equation (7)(ii) can be approximated by:

$$\begin{aligned}&\mathbb{E}_q[\log p(y \mid \theta)]\\ &= -\frac{1}{2\sigma^2}\mathbb{E}_q\left[\|y - g(\theta)\|_2^2\right]\\ &= -\frac{1}{2\sigma^2}\int q(z)\,\|y - g(z)\|_2^2\,dz\\ &\approx -\frac{1}{2\sigma^2}\sum_{\Theta^*\in\Phi,\Theta^*=(\theta^*,e^*)} q^*(\theta^*)\,\|y - g(\theta^*)\|_2^2\end{aligned} \tag{9}$$

where Φ is the collection of samples in the discretized sample space of (θ, e) and the generation of it was called the first-order sampling. In this work, we segmented the sample space with mean of posterior $\mu(\Theta)$ as the center and standard deviation of posterior as our resolution, i.e. for $\Theta = (\Theta_1, \Theta_2, \cdots, \Theta_m)$, $\Phi = \{(\Theta_1^*, \Theta_2^*, \cdots, \Theta_m^*) : \Theta_k^* \in \mu(\Theta_k) \pm n_1\sigma(\Theta_k), 1 \le k \le m\}$, where n_1 is a set of numbers determining the sample size. In the case of Gaussian posterior distribution, $n_1 = \{-3, -2, -1, 0, 1, 2, 3\}$ covers 99% of the whole distribution with resolution as 1σ and sample size 7. The probability $q(\theta)$ was normalized to $q^*(\theta)$ to guarantee $\sum_{\theta^*\in\Theta} q^*(\theta^*) = 1$.

In our study, the ground truth distributions of input signal y and recovered signal y' were both Gaussian distributions, KL loss between signal distributions shown in equation (7)(ii) can be simplified as:

$$\begin{aligned}&\mathrm{KL}\left[P(y)||P'\left(y'\right)\right]\\&=\frac{1}{2}\log\left(2\pi{\sigma'}^2\right)+\frac{\sigma^2+\left(\mu-\mu'\right)^2}{2{\sigma'}^2}-\frac{1}{2}\left(1+\log 2\pi\sigma^2\right)\\&=\log\frac{\sigma'}{\sigma}+\frac{\sigma^2+\left(\mu-\mu'\right)^2}{2{\sigma'}^2}\end{aligned}\tag{10}$$

where μ, σ denote the mean and standard deviation of the distribution P for input signal y and μ', σ' denote those of the distribution P' for recovered signal y'. During training process by using simulation data, the distribution of input signal is known, thus the only task is to evaluate μ' and σ', which requires the second sampling to generate the realizations of noise. From the first sampling, the sample space of noise parameter e has already been segmented, i.e. $e^* \in \{exp(\mu(log(e))\pm n_1\sigma(log(e)))\}$. However, this step only discretizes the sample space of e but the realizations of error ϵ requires another sampling from the noise distribution which is Guassian according to the assumption $\epsilon \sim N(0, e)$. Thus μ' and σ' can be written as:

$$\begin{aligned}\mu' &= \mathbb{E}_{P'}[y']\\&=\mathbb{E}_{P'}[g(\theta)+\epsilon]\\&=\int[g(\theta)+\epsilon]q(\theta)N(\epsilon)d\theta d\epsilon\\&\approx\sum_{\theta^*\in\Phi}\sum_{\epsilon^*\in E}[g(\theta^*)+\epsilon^*]q^*(\theta^*)N^*(\epsilon^*)\end{aligned}\tag{11}$$

$$\begin{aligned}{\sigma'}^2 &= \mathbb{E}_{P'}[{y'}^2]-{\mu_2}^2\\&=\mathbb{E}_{P'}[g(\theta)+\epsilon]^2-{\mu_2}^2\\&\approx\sum_{\theta^*\in\Phi}\sum_{\epsilon^*\in E}[g(\theta^*)+\epsilon^*]^2q^*(\theta^*)N^*(\epsilon^*)-{\mu_2}^2\end{aligned}\tag{12}$$

Here N is the probability density function of ϵ and $E = \{\epsilon^* : \epsilon^* = \pm n_2 e, e = exp(\mu(e) \pm n_1\sigma(e))\}$ is the collection of error realizations after the second sampling. Since the error term is assumed to be a Gaussian distribution, n_2 is also specified as $\{-3, -2, -1, 0, 1, 2, 3\}$ covering 99% of the entire distribution. To simplify the calculation, the independence of θ and ϵ is assumed. Similar to q^*, N^* is normalized after discretization.

IV. Experiments and Results

A. Forward Models for Model Evaluation

The framework was firstly evaluated using simulated data from a simple bi-exponential model:

$$y = Ae^{-at} + Be^{-bt}\tag{13}$$

where A, B were amplitude parameters and a, b were two rates. This provided a representative 'toy' model [6] to verify the feasiblity of the algorithm.

Subsequently the framework was tested using simulated data from two ASL models, the ASL General kinetic model and ASL gamma dispersion model, to test its performance on medical imaging models. These models reflect different levels of model complexity and are fairly representative of existing non-linear model fitting scenarios in quantitative MRI. The ASL model expressions can be summarised as:

$$y = M_0 fg(ATT, t_1, t_{1b}, \tau)\tag{14}$$

$$y = M_0 fg_{disp}(ATT, t_1, t_{1b}, \tau, s, p)\tag{15}$$

The ASL general kinetic model was described by Buxton [26] based on tracer kinetics to describe the relationship between time-series ASL data y and physiological and acquisition parameters, including perfusion f and arterial transit time (ATT), relaxation time of tissue t_1, relaxation time of blood t_{1b} and label duration τ. For this work, the expression of general KM can be written as the product of M_0, f and a non-linear generating function g, the complete expression of ASL kinetic models can be found in supplementary materials. M_0 is the arterial magnetization which requires a separate calibration step to estimate in ASL experiments.

Based on the general KM, the gamma dispersion KM incorporates then effects of dispersion into the model, where dispersion describes the variation in arrival time of blood through the vasculature due to fluid flow effects. Two more parameters, the sharpness s and time-to-peak p, within the model control the change in shape due to dispersion.

B. Training of VAE-like Framework

Neural networks were separately trained for each of the three different forward models. The details of parameters to estimate, known parameters and Residual Neural Network (ResNet) structure [27] in each case are summarized in table 1. For each model, two hundred thousand datasets were generated as training data. For the bi-exponential model, all known parameters and parameters of interest were randomly sampled from 0 to 3. For ASL models, the following ranges were considered: $f \in [0, 3]$ ml/g/min, $ATT \in [0, 3]$ s, $t_1 \in [0.8, 1.8]$ s, $t_{1b} \in [0.9, 2.4]$ s, $\tau \in [0.1, 3]$ s, $log(s) \in [1, 3]$, $log(p) \in [-3, -1]$ and M_0 was fixed to be 1500. All physiologically plausible values of parameters are covered in these ranges [25]. Gaussian white noise with signal noise ratio (SNR) in the range [1,1000] were randomly added to 75% of the data. During neural network training with ASL models, ASL signals were scaled by 100 and 200 for general KM and dispersion KM as data normalization. Independent Gaussian distributions were used as likelihood distribution and the posterior assumption for all parameters of interest during training process except the error term e. As the standard deviation of a Gaussian likelihood, the e, cannot be negative, the posterior for ϵ was considered as the Log-Normal distribution. A uniform distribution was used as the prior distribution for all of parameters. We employed a ResNet [27] with 100 neurons in each layer as the encoder whose weights were trained using Adam optimizer with initial learning rate = 0.001, β_1 = 0.9 , β_2 = 0.9. The value of λ was fixed to be 10 during training process.

Forward Models	Parameter to estimate	Known Parameter	ResNet Structure
Bi-exponential Model	$\theta = (a, B)$	(A, b)	20 Layers
General KM	$\theta = (M_0 f, ATT)$	(t_1, t_{1b}, τ)	40 Layers
Dispersion KM	$\theta = (M_0 f, ATT, s, p)$	(t_1, t_{1b}, τ)	40 Layers

TABLE I
The details of three forward models are shown here, where parameters of interests represent the parameters to estimate and known parameters indicate the models parameters known before inference.

C. Simulation Experiments

One thousand simulation data were generated from the bi-exponential and ASL General kinetic model and ASL gamma dispersion model respectively with amplitude B and rate a varying between 0 and 2 for bi-exponential example, perfusion and ATT varying between 0 and 200 ml/100g/min and between 0 and 2 seconds for ASL models. M_0 in ASL examples was treated as a known parameter with fixed value. Gaussian noise was added to generate data with SNR of infinity, 10, 5 and 2.5. MCMC and aVB were also used to estimate parameters from the simulation data, to compare the average computational time, the estimated values and the associated uncertainty to that of VAE-based method. The accuracy of inference framework was evaluated by calculating the error between ground truth and estimated values. The computational time was calculated as the average of 10 experiments. Each aVB optimization went through 100 iterations for all forward models while MCMC used 2000, 3000, 5000 burn-in epochs for bi-exponential model, ASL general KM and dispersion KM respectively and generated 500 samples for calculation in each case. The number of burn-in epochs was determined by performing MCMC multiple times with burn-in epochs ranging from 500 to 5000, at intervals of 500 epochs and using the smallest burn-in epochs that yielded the smallest error. All experiments were performed by using MacBook Pro with a 2.7 GHz Intel Core i5 processor.

Figure 2 shows the parameter estimation error from the simulation experiments across all methods and SNR. An overall consistency in estimated values was observed across the different inference frameworks on bi-exponential example and ASL forward models. Across all experiments, median error was zero or near to zero and large errors were associated with lower SNR (wider inter-quartile range). The VAE-like framework and MCMC performed well on perfusion and ATT estimation using both general KM and dispersion KM, but aVB exhibited a slightly larger error for the dispersion model inference. A greater error in perfusion and ATT was observed at large SNR levels when using the VAE-like structure as the inference framework than MCMC and aVB.

Figure 3 shows the estimated parameter uncertainty (95% credible interval of the marginal posterior) from the simulation experiments. As expected, for all methods parameter uncertainty increased with increased noise on the data. There was a tendency for the VAE to estimate a larger CI for amplitude parameters than MCMC but similar/smaller when it was a time/rate parameter. MCMC converged to a negligible CI when there was no noise, but for ASL forward models, the VAE-like networks did not.

Figure 4, 5 and 6 show the estimated 2D joint density at one specific parameter pair across different SNR using all three inference algorithms. For bi-exponential model, we set $a = 1$ and $B = 1$ as the representative parameter pair. For ASL KMs, typical perfusion and ATT values of healthy individuals, i.e. perfusion = 0.6 ml/g/min and ATT = 1.5s were used as the example. The joint density was calculated for aVB and VAE-like structure as multivariate Gaussian distribution and independent Gaussian distribution respectively, for MCMC the empirical density was computed by using the density of points that fell into each grid among 10,000 repetitive samples. The density plots from different methods shared similar location in parameter space, indicating a consistent range of parameter estimations. Different from MCMC where posterior distributions in principle converge to the true posterior, aVB and VAE-like model only approximates the posterior giving rise to a difference in the shape of the distribution in figure 4, 5 and 6 to that of MCMC.

Table 2 compares the computational time when using different inference algorithms and forward models, indicating a hundredfold to thousandfold improvement by using VAE-based method.

	VAE-like	aVB	MCMC
Bi-exponential	0.0195s	3.47s	4.43s
ASL General KM	0.0223s	4.39s	11.1s
ASL Dispersion KM	0.0224s	41.0s	81.9s

TABLE II
Computational time in Simulation Experiment: the computational time for each experiment was measured 10 times and took an average

	VAE-like	aVB
ASL General KM	2.709s	7.850s
ASL Dispersion KM	2.338s	242.506s

TABLE III
Computational time in Real-data Experiment: the computational time for each experiment was measured 10 times and took an average

D. Real Data Experiments

To test the practical application of VAE-based method on ASL data, an in-vivo pCASL dataset with 6 post label delays (PLDS) (0.25, 0.5, 0.75, 1.0, 1.25, 1.5) and resolution 64x64x24 taken from the example in the Oxford Neuroimaging primer [28] was considered, which was taken from the website https://users.fmrib.ox.ac.uk/~chappell/asl_primer/ex3/index.htm and used with permission. The original dataset contains 96 volumes and a separate calibration image, where

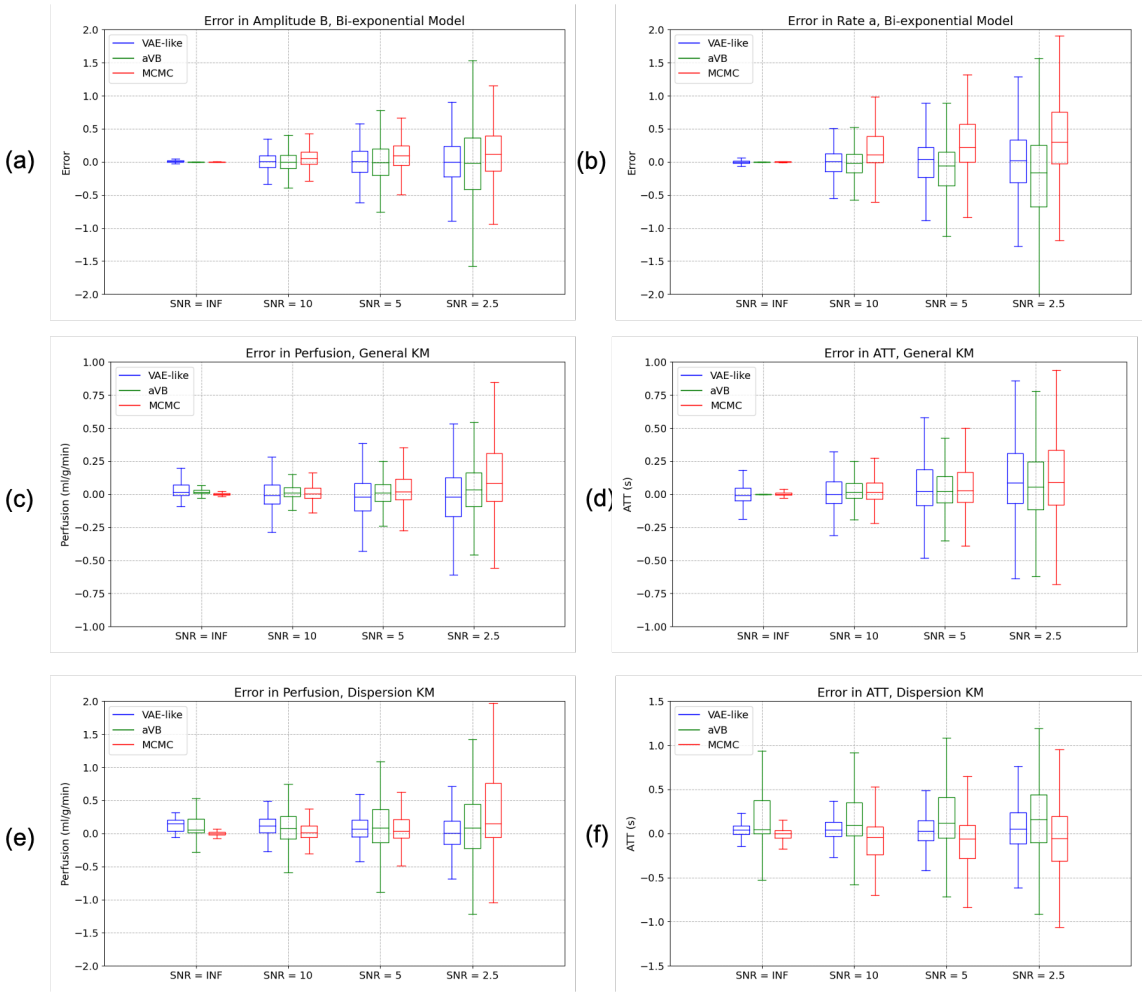

Fig. 2. Error of estimations across all three inference algorithms (MCMC, aVB and VAE-like framework): (a) and (b) error in amplitude B and rate a using bi-exponential model as the forward model;(c) and (d) error in perfusion and ATT using general KM as the forward model;(e) and (f) error in perfusion and ATT using dispersion KM as the forward model; The code for both simulation and real-data experiments in this work is available on GitHub https://github.com/Yechuan-z/VAE-like-framework.git.

each PLD was repeated 8 times. Before our experiment, pre-processing steps including motion correction and distortion correction were performed by using BASIL toolbox command *oxford asl* [29]. The total magnetization M_0 was also computed by this tool from the calibration image. An average of the ASL signal over 8 repeats and a reduced dataset containing 8 sets with a single repeat for each PLD were considered as the representative data at a high or low SNR levels. Perfusion and ATT values were measured by aVB and VAE-based methods by using general KM or dispersion KM as the forward model. As the brain image processing tools for ASL were already successfully implemented in the software FSL, the aVB analysis used in real data experiments was performed by BASIL toolbox command *basil* [29].

Figure 7 shows the estimated perfusion and ATT from the the averaged data when using general KM figure 7 (a) and dispersion KM figure 7 (b) as the forward model. In both cases, the perfusion parameter was successfully estimated by both aVB and VAE-like framework. A clear perfusion separation between white matter and grey matter can be observed which was consistent with the physiology of the brain. The VAE-like framework provided a longer ATT value in white matter, which was not as apparent from the ATT estimated using aVB.

Figure 9 in supplementary materials shows the estimated perfusion and ATT using a reduced data with only one measurement at each PLD by using general KM (a) and dispersion KM (b) as the forward model. Similar to figure 7, perfusion was estimated with a clear pattern by using both VAE-like framework and the aVB analysis, however perfusion estimations appeared to be less smooth by using VAE-like framework. Similar to averaged data set, VAE-like framework still exhibited a longer ATT value in white matter, but it was less noticeable when using aVB.

Table 3 shows the computational time when using aVB in BASIL tool box and using VAE-like framework to process the ASL pCASL dataset. The computational time for VAE-like framework stabilized at 2-3s for different choice of forward model, but aVB exhibited a 30-fold increase when using dispersion KM as forward model compared to that seen using the general KM.

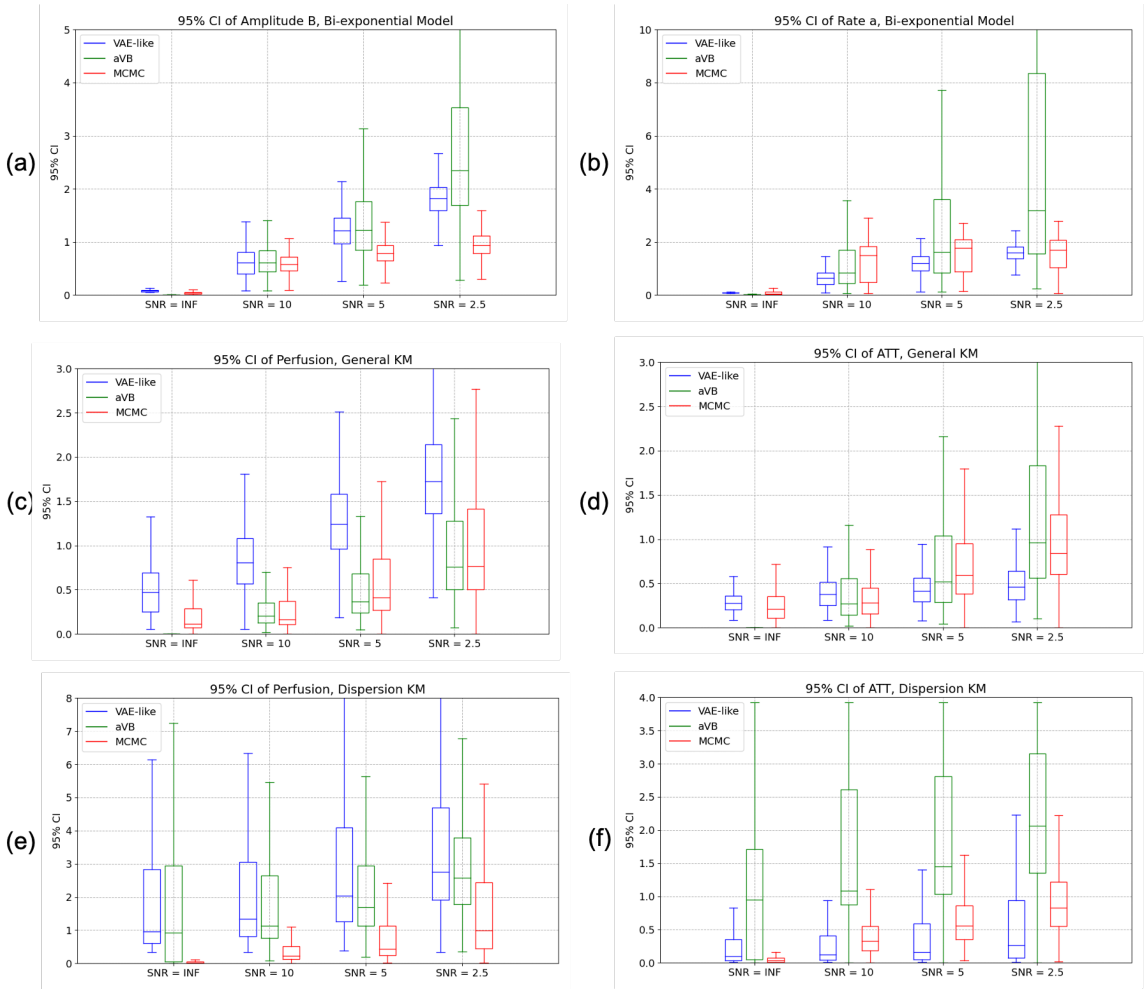

Fig. 3. 95% CI across all three inference algorithms (MCMC, aVB and VAE-like framework): (a) and (b) amplitude B and rate a using bi-exponential model as the forward model;(c) and (d) perfusion and ATT using general KM as the forward model;(e) and (f) perfusion and ATT using dispersion KM as the forward model;

V. Discussion

In this paper, a VAE-like framework was proposed to solve the parameter estimation problem for non-linear forward models, with a particular application in the field of medical imaging. Conventionally VAEs are used as a dimension compression technique, by adjusting the VAE structure and loss function it was adapted for parameter estimation via Bayesian inference. The framework was evaluated using a bi-exponential model and two ASL-MRI forward models on both synthetic data and real data. These methods represent relatively simple but typical forward models for data modelling with varying complexity, whilst still being amenable to computationally demanding Bayesian inference methods such as MCMC. MCMC and aVB were applied to compare with the VAE-like framework in simulation experiments, showing comparable results among different methods for parameter and uncertainty estimation. The number of MCMC burn-in epochs was determined by trial-and-error process as suggested in the previous section. Although this approach may not be the most efficient strategy to adopt, it should be sufficient to provide a reasonable comparison. Overall, the results from VAE-like frameworks were similar to 'gold standard' MCMC inference except a larger error for the VAE-like frameworks with ASL models at a low noise level. The observations of larger error at small noise levels might be explained by the fact that neural networks were trained with data simulated from all noise levels, whereas the estimation at a small noise level reflects the sum of all possible uncertainty of input signals covering all noise levels. This performance a low noise may be acceptable in practical applications such as ASL MRI which has an SNR within the lower range of those tested here.

The estimated 95% CI for each parameter of interest exhibited the expected increasing trend with noise but the magnitude varied among different inference methods. For ASL models, a greater CI on perfusion was observed when using the VAE-like framework than when using MCMC while that of ATT was similar between the two methods. The possible reason is that perfusion is a scaling factor of ASL signals, affecting the magnitude of the signal value, but ATT is an exponential term and also the boundary point of the piecewise ASL kinetic function, affecting the signal value more from the position of the peak rather than signal magnitude. Thus the model may be more likely to attribute the noise of the signal to the scaling

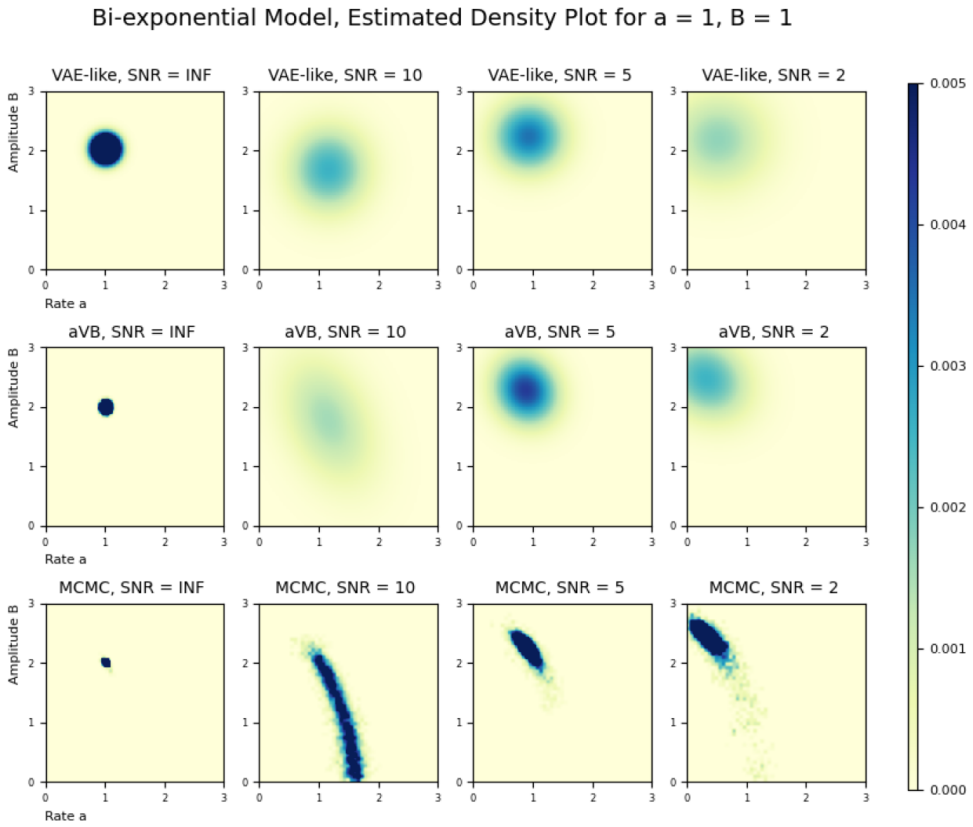

Fig. 4. 2D Plot of joint posterior for bi-exponential model with ground truth a = 1 and B = 1. Posterior distributions were generated by using VAE-like framework (first row), aVB (second row) and MCMC (third row). When performing MCMC for the bi-exponential model, 2000 burn-in epochs were considered.

factor perfusion than the boundary point ATT.

To test the similarity of posterior distribution among different inference algorithms including MCMC, aVB and VAE-like framework, the 2D density plots at some representative parameter values were given. In each case, aVB and the VAE-like framework were similar in both shape and location of the estimated approximate posterior. Although MCMC exhibited similar distribution location, the shapes of posterior were different from both aVB and the VAE-like framework. MCMC in principle can converge to the true posterior distribution, however, aVB and VAE-like framework follow the variational Bayesian approach of using an approximate posterior, which is restricted to a particular form, resulting in a natural deviation from the true distribution if the assumption of posterior is unable to exactly replicate the true form. From visual inspection, the shape of posterior estimated from aVB was more similar to that of MCMC than VAE-like framework. This may be explained by the fact that aVB assumed a multivariate Gaussian distribution with covariance terms but the current VAE-like framework assumed independent Gaussian posterior for each parameter. The additional covariance term allows more freedom in the choice of shape and is something that could in principle be added to a future implementation of the VAE-like framework.

We also tested the performance of VAE-like framework comparing to aVB on sample pCASL data, examining the performance when applied to higher SNR (average over multiple measurements) and subset of the full data. The results suggested VAE-like framework performed as well as aVB in perfusion inference when using general KM or dispersion KM as the forward model. The VAE-like frameworks also provided greater contrast between white and grey matter on ATT than aVB for both ASL models, which is consistent with the expected difference in ATT between these two tissue types.

In terms of calculation speed, the VAE-like framework was substantially faster for parameter inference. In contrast with aVB and MCMC, the VAE-like framework reduced computation cost by at least a hundred-fold in simulation experiments, enabling the computation of thousands of data by using ASL models within couple of seconds. While both MCMC and aVB required multiple of iterations to complete the parameter estimation process, the neural network amortized the repeated evaluations of forward model into the network training process. The improvement in computational time by using VAE-like framework in real data experiments was less substantial than simulation experiments. This may be explained by the overhead of loading and handling the imaging data which are included in the overall computation time. The computational cost for VAE-like framework was more stable among forward models of different complexity, but for aVB and MCMC a steep increase in computational time can be observed along with the model complexity. This has implications for the use of Bayesian inference in clinical applications of medical imaging, since the VAE-like approach might allow near real-

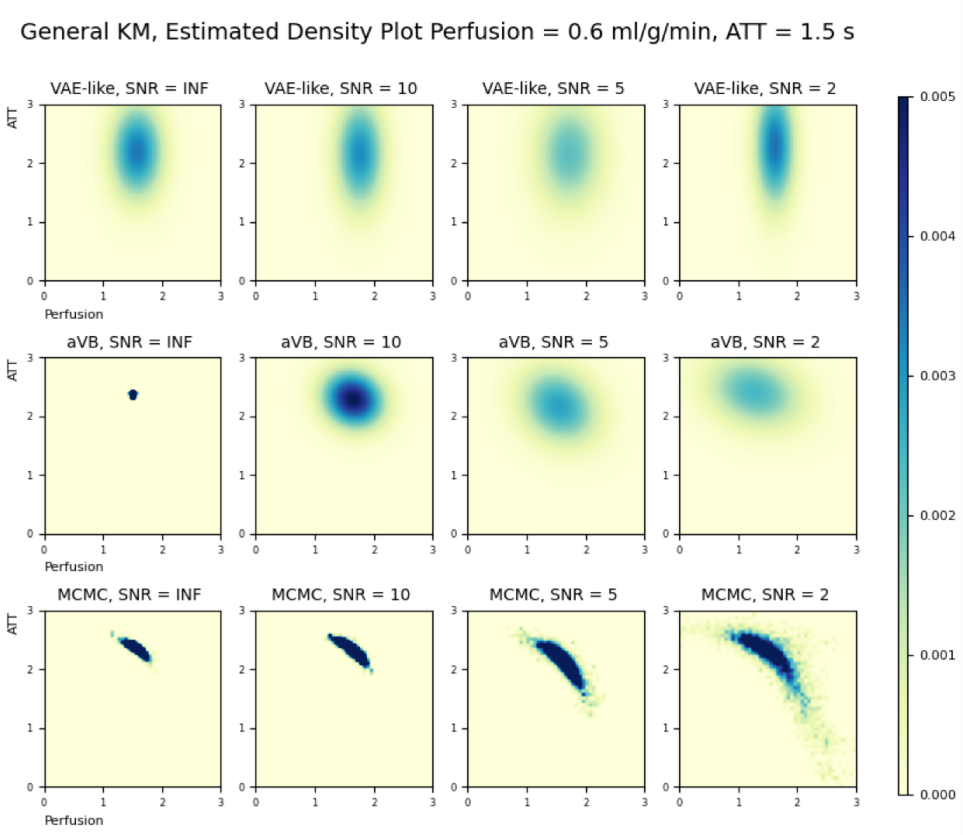

Fig. 5. 2D Plot of joint posterior for the ASL general KM with ground truth perfusion = 0.6 ml/g/min and ATT = 1.5s.Posterior distributions were generated by using VAE-like framework (first row), aVB (second row) and MCMC (third row). When performing MCMC for the ASL general KM, 3000 burn-in epochs were considered.

time production of parameter maps.

The proposed architecture of a VAE-like framework for parameter estimation is novel compared to other approaches to employ NNs within the context of medical image parameter mapping. Unlike previous literature which employed multi-layer perceptron [15] or CNN [17] to infer parameter of interest, the VAE-like structure adopted in this work allows for a complete (albeit approximate) Bayesian inference. There are some literature performing parameter inference by combining variational Bayes and neural network, for example, Bliesener et al. [22] in 2020 proposed a solution for DCE-MRI parameter inference, where the use of a variational loss enabled the calculation of parameter uncertainty, but they did not incorporate the forward model as the decoder and did not consider their work as an encoder-decoder pair. In our architecture, we constructed a complete VAE structure, where the encoder part is directly involved in parameter inference, and the decoder was used only in loss calculation during training. A novel loss calculation was established in the VAE-like framework through the incorporation of KL loss between distributions of signal and the incorporation of first-and-second order sampling strategy, while Bliesener's work only adopted the conventional loss.

Some previous literature have applied deep learning techniques to process image data, for example Luciw et al. [17] trained a CNN and a U-Net with real 3D data sets for automated perfusion estimation. But in our work neural networks were trained with simulation data, which avoids the possible bias from real data associated with different types of scanners, choices of imaging protocols and subject groups, and guarantees the generalizability to other data sets. The current neural network only takes vector data as an input, in principle this could be extended to image data in the future following the adoption of CNN in these other works whilst still retaining the ability to perform (approximate) Bayesian inference according to a defined forward model. A potential advantage of of adopting an image input would be having the ability to incorporate spatial information into the inference framework, mirroring the benefits seen when using a spatial prior in the aVB algorithm [30].

There are some potential limitations to the VAE-like framework. The most significant limitation is the specification of number of samples and the parameters of those samples. In designing this framework for ASL models, the choice of PLDs is fixed at beginning. However, in practice, the choice of PLDs can be flexible, for example, the use of optimal PLDs for estimating perfusion and ATT together in ASL-MRI [31]. For different acquisition protocols, a new VAE-like network needs to be established and trained before performing inference. This is not necessarily a limitation in practice as the computational complexity of training process depends on the application and the neural network training only needs to be done once

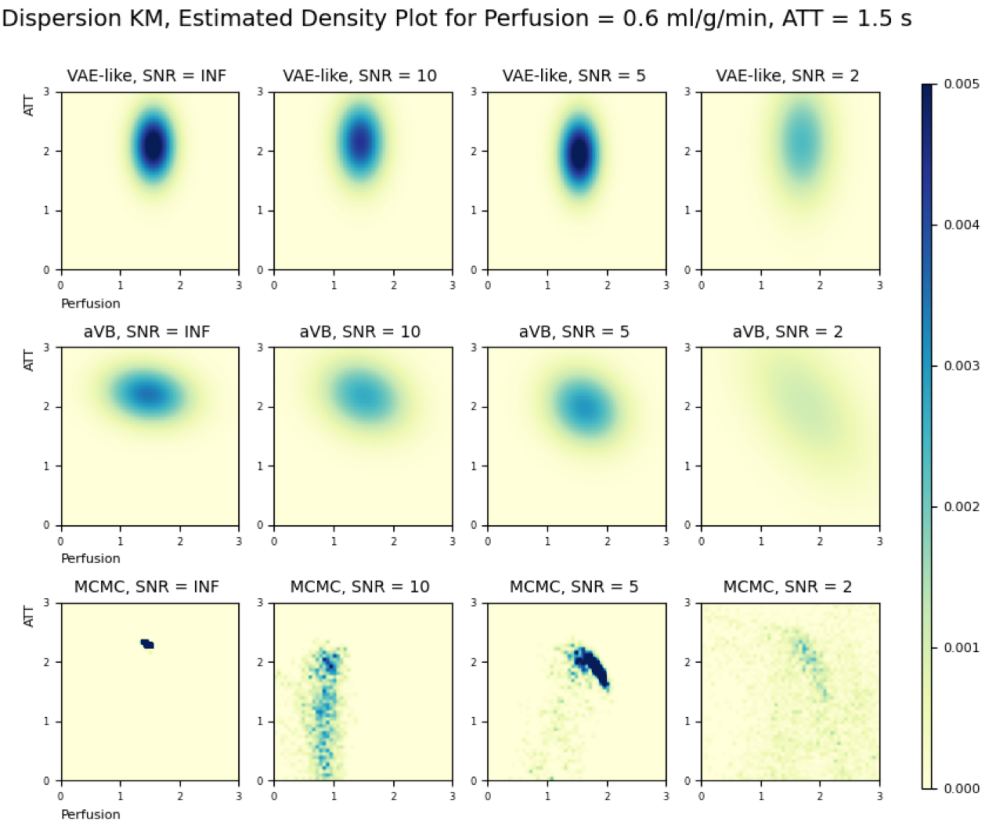

Fig. 6. 2D Plot of Joint Posterior for ASL dispersion KM with ground truth perfusion = 0.6 ml/g/min and ATT = 1.5s.Posterior distributions were generated by using VAE-like framework (first row), aVB (second row) and MCMC (third row). When performing MCMC for the ASL dispersion KM, 5000 burn-in epochs were considered.

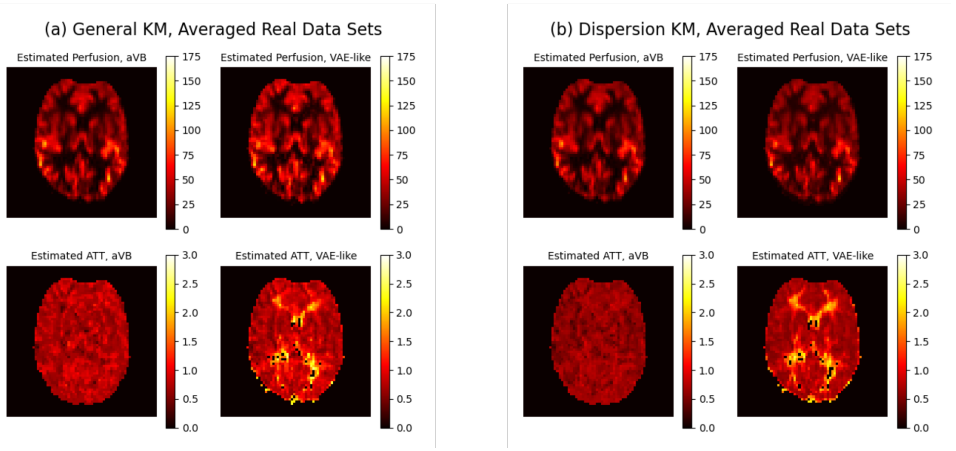

Fig. 7. Real Data Experiment with the Averaged Data Set: (a) Perfusion and ATT estimated by aVB and VAE-like structure by using the general KM as forward model; (b)Perfusion and ATT estimated by aVB and VAE-like structure by using the dispersion KM as forward model; The brain mask used for ATT values was generated from perfusion greater than 5 ml/100g/min.

according to the acquisition protocol and then can be reused for any data acquired with that protocol. One possible follow-up research direction would be including multiple choices of PLDs into one neural network.

The second limitation is a result of 'the curse of dimensionality'. There is a tendency of our VAE-like approach to become computationally expensive in generating samples during loss calculation in the training process as the number of parameters increases. Although the first-and-second order sampling strategy ensures the even distribution of samples, the number of samples required grows exponentially with number of latent parameters. For set n_1, n_2 and parameters θ, the number of samples to collect is $|n_1|^{|\theta|} * |n_2|$, where $|\cdot|$ denotes the cardinality. The only solution for this problem is to reduce either the range or the resolution of the sample collection, i.e. reducing $|n_1|$ or $|n_2|$, but at a cost of accuracy. Again, this

limitation only applies to the training of the inference network which would occur prior to deployment on new data when there is both potential for time and computational resources to be available.

The other obstacle in using such a framework is a potential lack of flexibility in the choosing the form of data likelihood, prior and posterior distribution. If the initial guess on prior or posterior changes, a new sampling strategy needs to be derived and a new neural network needs to be trained. More generally VB methods also face this problem, for example a change in prior distributional form would necessitate rederiving the update equations for aVB (if indeed it is possible to derive update equations at all).

VI. Conclusion

In this work, a fast VAE-based framework trained with a new loss function and a novel sampling strategy was proposed to solve parameter estimation problem in the context of nonlinear forward models with application to medical imaging. ASL kinetic models were used to test for the feasibility of this framework for parametric mapping. A hundredfold improvement was achieved in the computational efficiency compared with conventional inference frameworks without substantial loss of accuracy. The performance by using VAE-like framework on ASL data was as accurate as conventional methods in perfusion estimation and even outperformed in ATT estimation.

VII. Data Availability

In this work, an in-vivo pCASL dataset with 6 post label delays (PLDS) (0.25, 0.5, 0.75, 1.0, 1.25, 1.5) and resolution 64x64x24 was considered in real data experiments and used with permission. This dataset is openly available in Oxford Neuroimaging primer at website https://users.fmrib.ox.ac.uk/~chappell/asl_primer/ex3/index.htm.

VIII. Acknowledgement

This research was supported by EPSRC [grant number EP/P012361/1 and EP/P012361/2]. Jianqing Zheng is funded by Kennedy Trust Prize Studentships [grant number AZT00050-AZ04]. For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

References

[1] Green, Paul E. "Bayesian Decision Theory in Pricing Strategy." *Journal of Marketing*, vol. 27, no. 1, 1963, pp. 5–14. JSTOR, https://doi.org/10.2307/1248574.

[2] Coly, Sylvain et al. "Bayesian hierarchical models for disease mapping applied to contagious pathologies." *PloS one* vol. 16,1 e0222898. 13 Jan. 2021, doi:10.1371/journal.pone.0222898

[3] Patenaude, B. "Bayesian Statistical Models of Shape and Appearance for Subcortical Brain Segmentation." University of Oxford, 2007.

[4] Woolrich, Mark W et al. "Constrained linear basis sets for HRF modelling using Variational Bayes." *NeuroImage* vol. 21,4 (2004): 1748-61. doi:10.1016/j.neuroimage.2003.12.024

[5] Zhang, Linlin et al. "Bayesian Models for fMRI Data Analysis." *Wiley interdisciplinary reviews.* Computational statistics vol. 7,1 (2015): 21-41. doi:10.1002/wics.1339

[6] M. A. Chappell, A. R. Groves, B. Whitcher and M. W. Woolrich, "Variational Bayesian Inference for a Nonlinear Forward Model," in *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 223-236, Jan. 2009, doi: 10.1109/TSP.2008.2005752.

[7] Michael A. Chappell and Martin S. Craig and Mark W. Woolrich, "Stochastic Variational Bayesian Inference for a Nonlinear Forward Model", arXiv

[8] H. Attias, "A variational Bayesian framework for graphical models," in Proc. Adv. Neural Inf. Process. Syst., 2000.

[9] Chappell, M., et al. "The FMRIB Variational Bayes Tutorial: Variational Bayesian Inference for a Non-Linear Forward Model." Oxford Centre for Functional MRI of the Brain, 2016, pp. 1–23.

[10] L. Tierney and J. B. Kadane, "Accurate approximations for posterior moments and marginal densities," *J. Amer. Statist. Assoc.*, vol. 81, pp. 82–86, 1986.

[11] T. Leonard, J. S. J. Hsu, and K. W. Tsui, "Bayesian marginal inference," *J. Amer. Statist. Assoc.*, vol. 84, pp. 1051–1058, 1989.

[12] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, "Bayesian Data Analysis", 2nd ed. London, U.K.: Chapman and Hall/CRC, 2003.

[13] W.R.Gilks, S.Richardson, and D.Spiegelhalter, "Markov Chain Monte Carlo in Practice." London, U.K.: Chapman and Hall/CRC, 1996.

[14] Csiszar, I (February 1975). "I-Divergence Geometry of Probability Distributions and Minimization Problems". *Ann. Probab.* 3 (1): 146–158. doi:10.1214/aop/1176996454

[15] Lahiri, Anish et al. "Optimizing MRF-ASL scan design for precise quantification of brain hemodynamics using neural network regression." *Magnetic resonance in medicine* vol. 83,6 (2020): 1979-1991. doi:10.1002/mrm.28051

[16] Ma D, Gulani V, Seiberlich N, et al. "Magnetic resonance fingerprinting." *Nature*. 2013;495:187–192.

[17] Luciw, Nicholas J et al. "Automated generation of cerebral blood flow and arterial transit time maps from multiple delay arterial spin-labeled MRI." *Magnetic resonance in medicine* vol. 88,1 (2022): 406-417. doi:10.1002/mrm.29193

[18] Keiron O'Shea, Ryan Nash, "An Introduction to Convolutional Neural Networks", arXiv:1511.08458

[19] Olaf Ronneberger, Philipp Fischer, Thomas Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation", arXiv:1505.04597

[20] Diederik P Kingma, Max Welling, "Auto-Encoding Variational Bayes", arXiv:1312.6114, https://doi.org/10.48550/arXiv.1312.6114

[21] Daniel Jiwoong Im, Sungjin Ahn, Roland Memisevic, and Yoshua Bengio. 2017. "Denoising criterion for variational auto-encoding framework", arXiv:1511.06406

[22] Y. Bliesener, J. Acharya and K. S. Nayak, "Efficient DCE-MRI Parameter and Uncertainty Estimation Using a Neural Network," in *IEEE Transactions on Medical Imaging*, vol. 39, no. 5, pp. 1712-1723, May 2020, doi: 10.1109/TMI.2019.2953901.

[23] Williams, D S et al. "Magnetic resonance imaging of perfusion in the isolated rat heart using spin inversion of arterial water." *Magnetic resonance in medicine* vol. 30,3 (1993): 361-5. doi:10.1002/mrm.1910300314

[24] Detre, J.A., Leigh, J.S., Williams, D.S. and Koretsky, A.P. (1992), "Perfusion imaging." *Magn. Reson. Med.*, 23: 37-45. https://doi.org/10.1002/mrm.1910230106

[25] Grade, M et al. "A neuroradiologist's guide to arterial spin labeling MRI in clinical practice." *Neuroradiology* vol. 57,12 (2015): 1181-202. doi:10.1007/s00234-015-1571-z

[26] Buxton, R B et al. "A general kinetic model for quantitative perfusion imaging with arterial spin labeling." *Magnetic resonance in medicine* vol. 40,3 (1998): 383-96. doi:10.1002/mrm.1910400308

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", arXiv:1512.03385

[28] Mark Jenkinson, Michael Chappell, "Oxford Neuroimaging Primers", published by Oxford University Press.

[29] Michael Chappel, Martin Craig, "BASIL documentation"

[30] Groves, Adrian R et al. "Combined spatial and non-spatial prior for inference on MRI time-series." NeuroImage vol. 45,3 (2009): 795-809. doi:10.1016/j.neuroimage.2008.12.027

[31] Woods, Joseph G et al. "Designing and comparing optimized pseudo-continuous Arterial Spin Labeling protocols for measurement of cerebral blood flow." *NeuroImage* vol. 223 (2020): 117246. doi:10.1016/j.neuroimage.2020.117246

[32] Chappell, Michael A et al. "Modeling dispersion in arterial spin labeling: validation using dynamic angiographic measurements." *Magnetic resonance in medicine* vol. 69,2 (2013): 563-70. doi:10.1002/mrm.24260