School of Informatics

Report: Laplace prior and Variational Inference Probabilistic PCA algorithm (VIPPCA)

Xiaoyu Jiang
February 2023

1 Laplace distribution prior in Variational inference

1.1 Introduction

Compared to Gaussian distribution, the peak of Laplace distribution is sharper than Gaussian. This suggests that the number of samples of a Laplace source around zero is more than Gaussian [1]. The probability density function (PDF) of a one dimentional Laplace distribution has the form:

$$f(\theta) = \frac{1}{2b}\exp\left(-\frac{|\theta-\mu|}{b}\right)$$

where mean is μ and variance is $2b^2$. If we fix $\mu = 0$, we have:

$$f(\theta) = \frac{1}{2b}\exp\left(-\frac{|\theta|}{b}\right)$$

Therefore, for $\theta_1, \theta_2, ..., \theta_D$ which are independent and Laplace distributed, we have joint multivariate PDF as

$$p(\theta_1, \theta_2, \cdots \theta_D) = \frac{1}{(2b)^D}\exp\left(-\sum_{i=1}^{D}\frac{|\theta_i|}{b}\right)$$

1.2 ELBO Gradients with Laplace prior

Consider factor analysis variational distribution

$$q(\theta) = \mathcal{N}\left(\theta \mid \mathbf{c}, FF^T + \Psi\right)$$

where $\theta, \mathbf{c} \in \mathbb{R}^D$, $F \in \mathbb{R}^{D\times K}$ and $\Psi \in \mathbb{R}^{D\times D}$; $D \gg K$.

Same as (5.1) in MSc thesis, we need to compute

$$\nabla_{\mathbf{c},\mathbf{F},\Psi}\mathbb{E}_{q(\theta)}[\log q(\theta)] - \nabla_{\mathbf{c},\mathbf{F},\Psi}\mathbb{E}_{q(\theta)}[\log p(\theta)] - \nabla_{\mathbf{c},\mathbf{F},\Psi}\mathbb{E}_{q(\theta)}[\log p(\mathscr{D} \mid \theta)]$$

In particular, we focus on the second term, as the first and last term have no change.

Consider Laplace prior $p(\theta)$ with each single variable independent with each other:

$$p(\theta) = p(\theta_1, \theta_2, \cdots \theta_D) = \frac{1}{(2b)^D}\exp\left(-\sum_{i=1}^{D}\frac{|\theta_i|}{b}\right)$$

$$\log p(\theta_1, \cdots \theta_D) = \text{ constant } -\sum_{i=1}^{D}\frac{|\theta_i|}{b}$$

Then, the expectation of log likelihood would be:

$$\begin{aligned}\mathbb{E}_{q(\theta)}[\log p(\theta)] &= \text{ constant } -\frac{1}{b}\cdot\mathbb{E}_{q(\theta)}\left[\sum_{i=1}^{D}|\theta_i|\right] \\ &= \text{ constant } -\frac{1}{b}\sum_{i=1}^{D}\mathbb{E}_{q(\theta_i)}\left[|\theta_i|\right]\end{aligned}$$

denote $y_i = \mathbb{E}_{q(\theta_i)}[|\theta_i|]$, therefore we now consider how to compute $y = y_1 + y_2 + ... + y_D$. we denote that

$$F = \begin{bmatrix} -r_1^\top - \\ -r_2^\top - \\ \vdots \\ -r_D^\top - \end{bmatrix} \quad F^\top = \begin{bmatrix} | & | & & | \\ r_1 & r_2 & \cdots & r_D \\ | & | & & | \end{bmatrix} \quad r_i \in \mathbb{R}^K$$

$$\text{Thus,} \quad q(\theta_i) = \mathcal{N}\left(\theta_i \mid c_i, r_i^\top r_i + \psi_i\right) \quad \psi_i = \Psi_{ii}$$

Theorem: $\mathbb{E}[|X|] = \mu\left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right] + \frac{2\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}$ for $X \sim \mathcal{N}(\mu, \sigma^2)$ where Φ is the CDF function of standard normal distribution. (Proof see below 1.3)

By using this theorem, plug in $\mu = c_i$, $\sigma^2 = r_i^T r_i + \psi_i$, we have y_i:

$$y_i = \mathbb{E}_{q(\theta_i)}[|\theta_i|] = c_i \cdot \left[2\Phi\left(\frac{c_i}{\sqrt{r_i^\top r_i + \psi_i}}\right) - 1\right] + \frac{2\sqrt{r_i^\top r_i + \psi_i}}{\sqrt{2\pi}} . \exp\left(-\frac{c_i^2}{2\left(r_i^\top r_i + \psi_i\right)}\right)$$

where the Φ refers to the CDF function of standard normal distribution.

Introducing notation $z_i = r_i^\top r_i + \psi_i$, we rewrite y_i as:

$$y_i = \mathbb{E}_{q(\theta_i)}[|\theta_i|] = c_i \cdot \left[2\Phi\left(\frac{c_i}{\sqrt{z_i}}\right) - 1\right] + \frac{2\sqrt{z_i}}{\sqrt{2\pi}} . \exp\left(-\frac{c_i^2}{2(z_i)}\right)$$

We use the function g denotes the probability density function of standard normal, that is $g(x) = \Phi'(x) = \mathcal{N}(x \mid 0, 1)$

$$\begin{aligned}
\frac{\partial y_i}{\partial c_i} &= 2\Phi\left(\frac{c_i}{\sqrt{z_i}}\right) - 1 + 2c_i \cdot \frac{1}{\sqrt{z_i}} g\left(\frac{c_i}{\sqrt{z_i}}\right) + \sqrt{\frac{2}{\pi}}\sqrt{z_i} \exp\left(-\frac{c_i^2}{2z_i}\right)\left(-\frac{c_i}{z_i}\right) \\
&= 2\Phi\left(\frac{c_i}{\sqrt{z_i}}\right) - 1 + 2 \cdot \frac{c_i}{\sqrt{z_i}} \mathcal{N}\left(\frac{c_i}{\sqrt{z_i}} \mid 0, 1\right) - \sqrt{\frac{2}{\pi}} \cdot \frac{c_i}{\sqrt{z_i}} \cdot \exp\left(-\frac{c_i^2}{2z_i}\right) \\
&= 2\Phi\left(\frac{c_i}{\sqrt{z_i}}\right) - 1
\end{aligned}$$

$$\begin{aligned}
\frac{\partial y_i}{\partial z_i} &= -2 \cdot c_i g\left(\frac{c_i}{\sqrt{z_i}}\right) c_i \cdot z_i^{-\frac{3}{2}} \cdot \frac{1}{2} + \sqrt{\frac{2}{\pi}} \left\{\frac{1}{2\sqrt{z_i}} \exp\left(-\frac{c_i^2}{2z_i}\right) + \sqrt{z_i} \cdot \exp\left(-\frac{c_i^2}{2z_i}\right) \cdot \frac{c_i^2}{2z_i^2}\right\} \\
&= -c_i^2 N\left(\frac{c_i}{\sqrt{z_i}} \mid 0, 1\right) z_i^{-\frac{3}{2}} + \frac{1}{\sqrt{2\pi}} \left\{z_i^{-\frac{1}{2}} + c_i^2 z_i^{-\frac{3}{2}}\right\} \exp\left(-\frac{c_i^2}{2z_i}\right) \\
&= -c_i^2 N\left(\frac{c_i}{\sqrt{z_i}} \mid 0, 1\right) z_i^{-\frac{3}{2}} + \frac{1}{\sqrt{2\pi}} \left\{z_i^{-\frac{1}{2}}\right\} \exp\left(-\frac{c_i^2}{2z_i}\right) + \frac{1}{\sqrt{2\pi}} c_i^2 z_i^{-\frac{3}{2}} \exp\left(-\frac{c_i^2}{2z_i}\right) \\
&= \frac{1}{\sqrt{2\pi}} \left\{z_i^{-\frac{1}{2}}\right\} \exp\left(-\frac{c_i^2}{2z_i}\right)
\end{aligned}$$

By chain rule, we have $\frac{\partial y_i}{\partial \psi_i} = \frac{\partial y_i}{\partial z_i}$

$$\begin{aligned} dy_i &= \mathrm{tr}\left(\frac{\partial y_i}{\partial z_i} dz_i\right) = \mathrm{tr}\left(\frac{\partial y_i}{\partial z_i}\left[(dr_i)^\top r_i + r_i^\top dr_i\right]\right) \\ &= \mathrm{tr}\left(\frac{\partial y_i}{\partial z_i}(dr_i)^\top r_i\right) + \mathrm{tr}\left(\frac{\partial y_i}{\partial z_i} r_i^\top dr_i\right) \\ &= \mathrm{tr}\left(2\frac{\partial y_i}{\partial z_i} r_i^\top dr_i\right) \\ &= \mathrm{tr}\left\{\left[2 \cdot \frac{\partial y_i}{\partial z_i} \cdot r_i\right]^\top dr_i\right\} \end{aligned}$$

Therefore, $\nabla_{r_i} y_i = 2 \cdot \frac{\partial y_i}{\partial z_i} \cdot r_i$

We then consider $y = y_1 + y_2 + ... + y_D$, thus the objective $\mathbb{E}_{q(\theta)}[\log p(\theta)] = \text{ constant } - \frac{1}{b} \cdot y$ and we write derivatives in vector and matrix form: $\mathbf{c}, \mathbf{z} = diag(FF^T + \Psi), \psi$ are the vector collection of previous D terms.

$$\nabla_{\mathbf{c}} y = 2\Phi\left(\frac{\mathbf{c}}{\sqrt{\mathbf{z}}}\right) - 1 \tag{1}$$

$$\nabla_{\mathbf{z}} y = \nabla_{\psi} y = \frac{1}{\sqrt{2\pi}} \mathbf{z}^{-\frac{1}{2}} \odot \exp\left(-\frac{\mathbf{c}^2}{2\mathbf{z}}\right) \tag{2}$$

$$\nabla_{F^T} y = 2 \cdot F^T \cdot diagonalize(\nabla_{\mathbf{z}} y) \tag{3}$$

$$\nabla_{F} y = 2 \cdot diagonalize(\nabla_{\mathbf{z}} y) \cdot F \tag{4}$$

Note that all operation for vectors in (1) and (2) are all element-wise. The *diagonalize* here refers to transforming a vector to a diagonal matrix by putting elements on the matrix diagonal.

The re-parameterisation $\psi = \exp(\gamma)$ can be used to ensure that the variances remain positive. Since $\nabla_\gamma \psi = \psi$,

$$\nabla_\gamma y = \nabla_\psi y \odot \nabla_\gamma \psi = \frac{1}{\sqrt{2\pi}} \mathbf{z}^{-\frac{1}{2}} \odot \exp\left(-\frac{\mathbf{c}^2}{2\mathbf{z}}\right) \odot \psi \tag{5}$$

To get the final gradients for parameter update, multiply (1), (4) and (5) by $-\frac{1}{b}$

1.3 Practical implementation

By rewritting $\frac{1}{b}$ as α (re-using notations here, different meaning to the one used in the MSc thesis). So here $\alpha = \sqrt{\frac{2}{V}}$, V means the variance of Laplace distribution. The practical implementation would be:

$$\mathbf{c} \leftarrow \mathbf{c} - \eta_{\mathbf{c}} \left(\alpha \cdot [2\Phi\left(\frac{\mathbf{c}}{\sqrt{\mathbf{z}}}\right) - 1] + \frac{1}{L}\sum_{l=1}^{L} N \mathbf{g}_{\theta_l}\right) \tag{6}$$

$$\mathbf{F} \leftarrow \mathbf{F} - \eta_{\mathbf{F}} \left(-\mathbf{A} + \mathbf{C}\mathbf{B}^{\top} + 2\alpha \cdot diagonalize[\frac{1}{\sqrt{2\pi}} \mathbf{z}^{-\frac{1}{2}} \odot \exp\left(-\frac{\mathbf{c}^2}{2\mathbf{z}}\right)] \cdot F + \frac{1}{L} \sum_{l=1}^{L} N \mathbf{g}_{\theta_l} \mathbf{h}_l^{\top} \right) \tag{7}$$

$$\gamma \leftarrow \gamma - \eta_{\gamma} \left(-\frac{1}{2} + \frac{1}{2} \text{sum}(\mathbf{C} \odot \mathbf{A}, \text{dim} = 1) \odot \psi + \alpha \cdot \frac{1}{\sqrt{2\pi}} \mathbf{z}^{-\frac{1}{2}} \odot \exp\left(-\frac{\mathbf{c}^2}{2\mathbf{z}}\right) \odot \psi + \frac{1}{L} \sum_{l=1}^{L} \frac{N}{2} \mathbf{g}_{\theta_l} \odot \left(\psi^{1/2} \odot \mathbf{z}_l\right) \right) \tag{8}$$

Here the black terms are same as equation (5.28) - (5.30) in MSc thesis, red terms are the special terms for Laplace prior.
There are some practical issue need to be considered. First issue is the computation of $\mathbf{z} \in \mathbb{R}^D$. By definition, $\mathbf{z}$ are taken from the diagonal of $D \times D$ matrix $FF^T + \Psi$, for which naive implementation is computationally expensive. Practically we use

$$\mathbf{z} = diag(FF^T + \Psi) = diag(FF^T) + diag(\Psi)$$
$$\text{where} \quad diag(FF^T) = sum(F \odot F, dim = 1)$$

Another practical concern is the matrix multiplication in 7. Since the diagonalize operation will return a D by D matrix, we should avoid this. The approach to do this is via broadcasting:

$$\mathbf{F} \leftarrow \mathbf{F} - \eta_{\mathbf{F}} \left(-\mathbf{A} + \mathbf{C}\mathbf{B}^{\top} + 2\alpha \cdot \frac{1}{\sqrt{2\pi}} \mathbf{z}^{-\frac{1}{2}} \odot \exp\left(-\frac{\mathbf{c}^2}{2\mathbf{z}}\right) \odot F + \frac{1}{L} \sum_{l=1}^{L} N \mathbf{g}_{\theta_l} \mathbf{h}_l^{\top} \right) \tag{9}$$

Proof 1

For Gaussian random variable $X \in \mathbb{R}$ follows $\mathcal{N}(X \mid \mu, \sigma^2)$:

$$\begin{aligned}
\mathbb{E}[|X|] &= \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} dx \\
&= \left(y = \frac{x-\mu}{\sigma}\right) \int_{-\infty}^{\infty} |\mu + \sigma y| \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \\
&= \int_{-\frac{\mu}{\sigma}}^{\infty} (\mu + \sigma y) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy - \int_{-\infty}^{-\frac{\mu}{\sigma}} (\mu + \sigma y) \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \\
&= \mu \left(\int_{-\frac{\mu}{\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy - \int_{-\infty}^{-\frac{\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{y^2}{2}\right\} dy \right) + \\
&\quad \frac{\sigma}{\sqrt{2\pi}} \left(\int_{-\frac{\mu}{\sigma}}^{\infty} y \exp\left\{-\frac{y^2}{2}\right\} dy - \int_{-\infty}^{-\frac{\mu}{\sigma}} y \exp\left\{-\frac{y^2}{2}\right\} dy \right) \\
&= \mu \left[\Phi\left(\frac{\mu}{\sigma}\right) - \Phi\left(-\frac{\mu}{\sigma}\right) \right] + \frac{\sigma}{\sqrt{2\pi}} \left[\exp\left\{-\frac{y^2}{2}\right\} \Big|_{\infty}^{-\frac{\mu}{\sigma}} - \exp\left\{-\frac{y^2}{2}\right\} \Big|_{-\frac{\mu}{\sigma}}^{-\infty} \right] \\
&= \mu \left[2\Phi\left(\frac{\mu}{\sigma}\right) - 1 \right] + \frac{2\sigma}{\sqrt{2\pi}} \exp\left\{-\frac{\mu^2}{2\sigma^2}\right\}
\end{aligned}$$

Algorithm 1 VIFA – Laplace Prior

Input: Dataset $\mathscr{D} = \{(x_n, y_n)\}_{n=1}^N$, observation dimension D, latent dimension K, $\alpha > 0$, iterations T, mini-batch size M, Monte Carlo average size L, learning rates $\eta_c, \eta_F, \eta_\gamma > 0$

1: Initialise $c = 0^D$, $F \in^{D\times K}$, $\psi = 1^D$
2: $\gamma \leftarrow \log \psi$
3: Initialise $\hat{g}_c = 0^D$, $\hat{G}_F = 0^{D\times K}$, $\hat{g}_\gamma = 0^D$
4: $A \leftarrow \psi^{-1} \odot F$ (with broadcasting)
5: $B \leftarrow F^A$
6: $C \leftarrow A(I+B)^{-1}$
7: $\mathbf{zz} \leftarrow \psi + \text{sum}(F \odot F,\ \text{dim} = 1)$
8: $\mathbf{ww} \leftarrow \frac{1}{\sqrt{2\pi}} \mathbf{zz}^{-\frac{1}{2}} \odot \exp\left(-\frac{c^2}{2\mathbf{zz}}\right)$
9: **for** $t = 1, \ldots, T$ **do**
10: Sample a mini-batch $\mathscr{B}$ of M training examples from $\mathscr{D}$
11: Sample $h \sim \mathscr{N}(0^K, I^{K\times K})$
12: Sample $z \sim \mathscr{N}(0^D, I^{D\times D})$
13: $\theta \leftarrow Fh + c + \psi^{1/2} \odot z$
14: $g_\theta \leftarrow \nabla_\theta \big(-\frac{1}{M} \sum_{(x,y)\in\mathscr{B}} \log p(y|x,\theta) \big)$ (via back-propagation)
15: $\hat{g}_c \leftarrow \hat{g}_c + N g_\theta$
16: $\hat{G}_F \leftarrow \hat{G}_F + N g_\theta h$
17: $\hat{g}_\gamma \leftarrow \hat{g}_\gamma + \frac{N}{2} g_\theta \odot \big(\psi^{1/2} \odot z\big)$
18: **if** $t \mod L = 0$ **then**
19: $c \leftarrow c - \eta_c \big(\alpha \cdot [2\Phi\left(\frac{c}{\sqrt{\mathbf{zz}}}\right) - 1] + \frac{1}{L}\hat{g}_c \big)$
20: $F \leftarrow F - \eta_F \big(-A + CB^T + 2\alpha \cdot \mathbf{ww} \odot F + \frac{1}{L}\hat{G}_F \big)$
21: $\gamma \leftarrow \gamma - \eta_\gamma \big(-\frac{1}{2} + \frac{1}{2}\text{sum}(C \odot A,\ \text{dim} = 1) \odot \psi + \alpha \cdot \mathbf{ww} \odot \psi + \frac{1}{L}\hat{g}_\gamma \big)$
22: $\psi \leftarrow \exp \gamma$
23: Reset $\hat{g}_c, \hat{G}_F$ and $\hat{g}_\gamma$ according to line 3
24: Recalculate A, B, C, $\mathbf{zz}$ and $\mathbf{ww}$ according to lines 4-8
25: **end if**
26: **end for**
27: **return** c, F, ψ

2 Probabilistic PCA modification for VIFA algorithm: VIPPCA

Experimental results of VIFA algorithms with larger neural networks show that factor analysis variational distribution is prone to overfit, which means larger model will have worse performance. This might due to factor analysis distribution is over-flexible. We restrict the per-dimensional noise by replacing factor analysis distribution by Probabilistic PCA. The diagonal matrix Ψ no longer permits a different variance for each dimension, but all have the same value σ, in other words, variational distribution now becomes $\mathcal{N}(\theta \mid \mathbf{c}, FF^T + \sigma \cdot \mathbb{I})$. σ is regraded as a hyper-parameter tuned via random search sampled log-uniformly from $[0.01, 1]$. We call the new algorithm with probabilistic PCA as variational distribution as VIPPCA algorithm, compared to VIFA algorithm with factor analysis as variational distribution.

3 Experimental results

3.1 Laplace Prior

To make experimental results comparable to that of the Gaussian prior, we make sure the Laplace prior has the same variance as the variance used for Gaussian prior in MSc thesis. Note that in MSc thesis, α refers to prior precision, but in this report α refers to $1/b$, and variance of Laplace prior is $2b^2$. Therefore, the α parameter for laplace prior is set to be $\sqrt{2 \times \textit{precision}}$, where *precision* is value of α used in Gaussian prior experiments.

3.1.1 Posterior Estimation with Synthetic Data

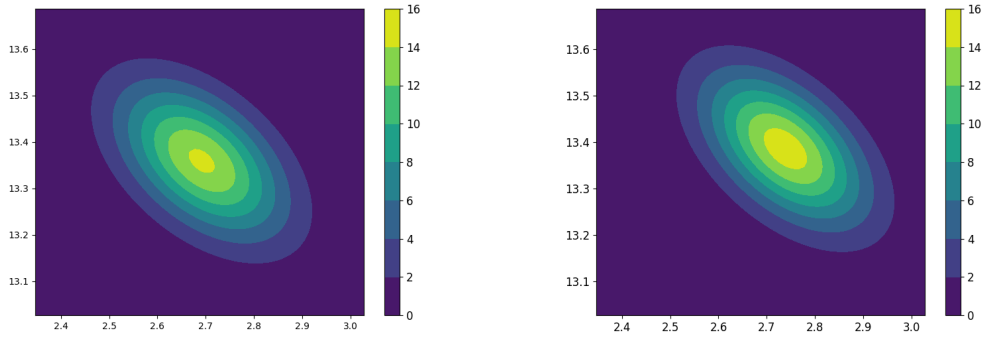
We first test the effectiveness of VIFA with laplace prior for linear regression on a simple synthetic dataset. All settings are same as the ones in MSc thesis, except now α for laplace prior is now $\sqrt{2 \times \textit{precision}}$, where *precision* is the chosen value to be set for the precision of the prior distribution.

The aggregated results of applying VIFA to the data are shown in Table 1. A qualitative comparison between the true and approximate posteriors in also given in Figure 1 for one of the 10 trials.

	Rel. Dist. from Mean	Rel. Dist. from Covariance	Scaled 2-Wasserstein Dist.
Gaussian Prior	0.0031±0.0005	0.0983±0.0129	0.0194±0.0023
Laplace Prior	0.0034±0.0006	0.0964±0.0113	0.0235±0.0035

Table 1: Distances between the true posterior distribution of the parameter vector of a linear regression model and the approximate FA posterior estimated using VIFA. The results for Gaussian Prior is directly taken from MSc thesis.

These results shows that VIFA with laplace prior can also do a good job of approximating the true posterior covariance for this synthetic setting, and shows similar performance to VIFA with gaussian prior considering table 1 .



(a) Ground truth

(b) VIFA-Laplace

Figure 1: The true posterior pdf of a linear regression model with two learnable parameters, plus the pdf of a FA model with a single latent dimension which was fit to the same data using VIFA with Laplace prior.

3.1.2 Posterior Estimation on UCI Datasets

We further explore the effectiveness of VIFA-Laplace for linear regression on UCI datasets. Same hyper-parameters are used for Laplace prior on UCI datasets. Same as MSc thesis, `BayesianRidge` automatically selects *precision* via the Bayesian method, in Laplace prior case we rescale it via $\alpha = \sqrt{2} \times precision$.

Figures 3, 4, 5 and 2 show a qualitative comparison of the true posterior of a linear regression model and the approximate posterior learned by VIFA (with Laplace prior), in the case of the Boston Housing, Concrete Strength, Energy Efficiency and Yacht Hydrodynamics datasets, respectively.

One observed fact is that compared to the results for gaussian prior, the gap between ground truth posterior variance (covariance) and approximated posterior variance (covariance) is larger when laplace prior is in use. For instance, in Figure 2(b) the approximated posterior variances for many features are less than one half of the true posterior variances, but this does not happen for gaussian prior case in Figure 5.2 (b) in MSc thesis.

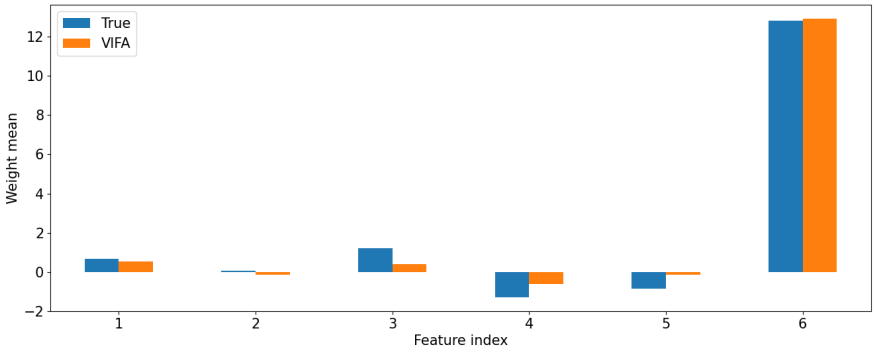
3.1.3 Neural Network Predictions

For neural networks, prior precision is a hyper-parameter, which is tuned via random search. In MSc thesis, precision for gaussian prior is log-uniformaly sampled from [0.01, 10]. To ensure the range of precision of laplace prior is also [0.01, 10], we choose α to be log-uniformaly sampled from $[\sqrt{2 \times 0.01}, \sqrt{2 \times 10}] \approx$[0.141, 4.472] [1].

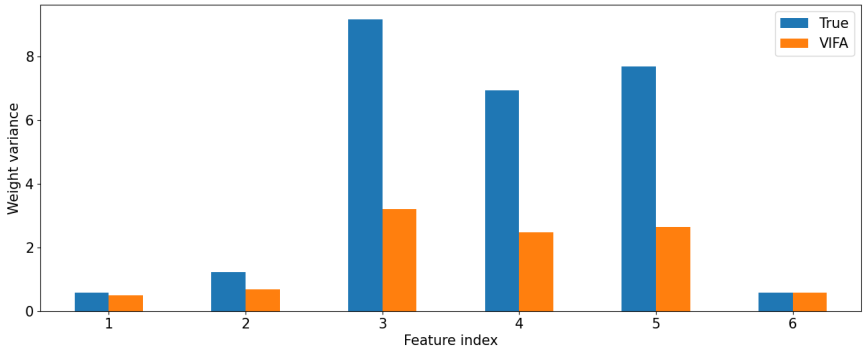
Other settings as same as the ones in MSc thesis, but we denote this single hidden layer 50 neurons network as SmallNN in the table 2, as we have experiments with larger neural networks in the next section.

The results shows that for neural network predictions, VIFA with laplace prior will leads to similar performance to that of gaussian prior. Laplace prior slightly outperform Gaussian prior

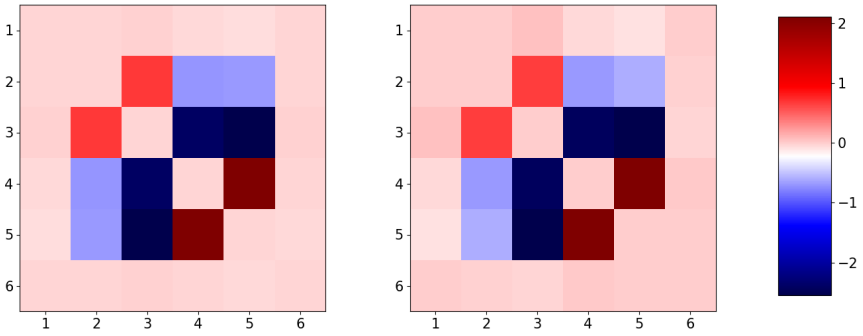
[1] In appendix, there are also results for $\alpha \in [0.01, 10]$



(a) Comparison of true and estimated posterior means

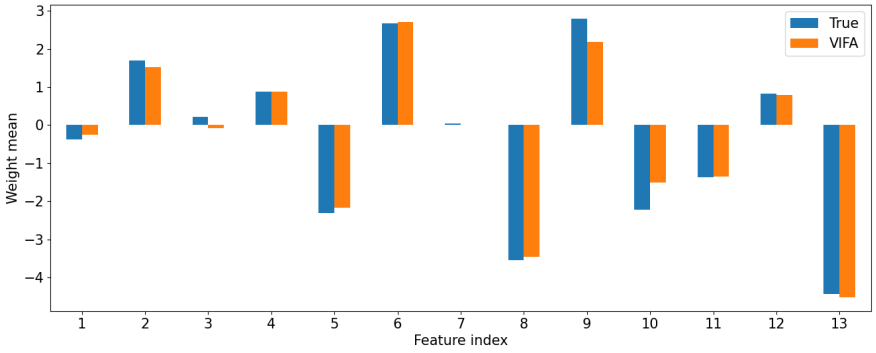


(b) Comparison of true and estimated posterior variances

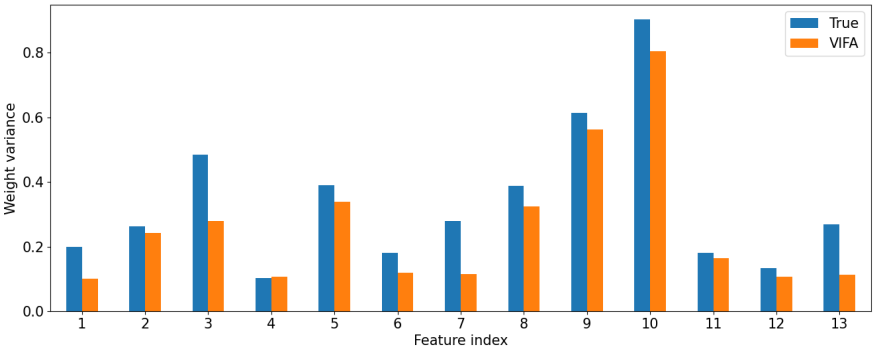


(c) Comparison of true and estimated posterior covariances

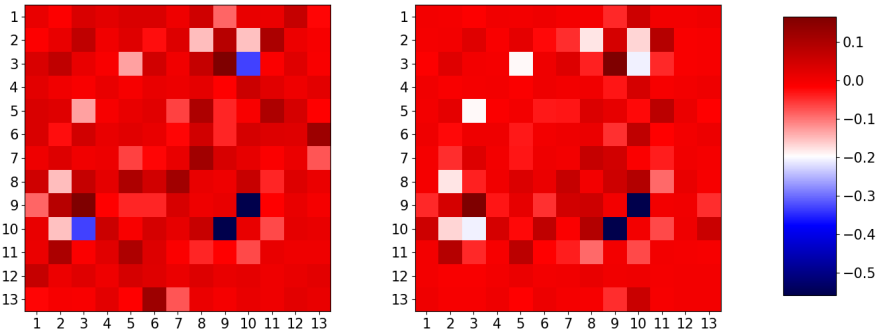
Figure 2: Comparison of the ground truth posterior of a linear regression model fit to the Yacht Hydrodynamics dataset, and the approximate posterior learned by VIFA (with Laplace prior). Variances and covariance are plotted separately due to the difference in their magnitude. In plot (c), the diagonal entries of the covariance matrices have been set to zero.



(a) Comparison of true and estimated posterior means

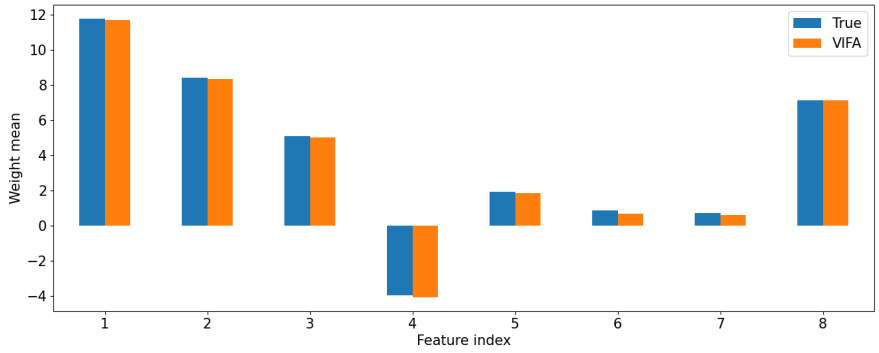


(b) Comparison of true and estimated posterior variances

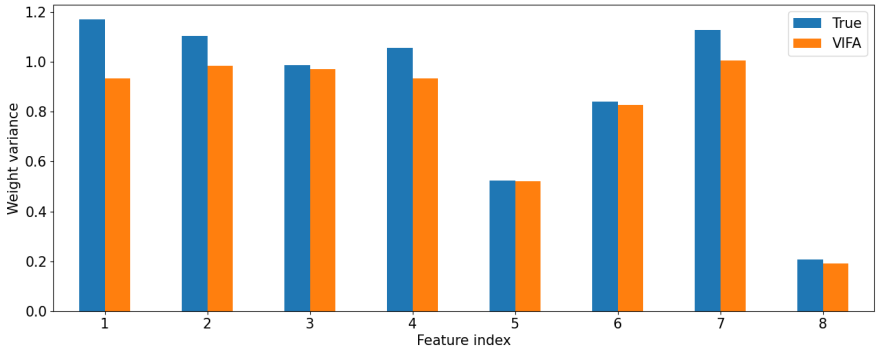


(c) Comparison of true and estimated posterior covariances

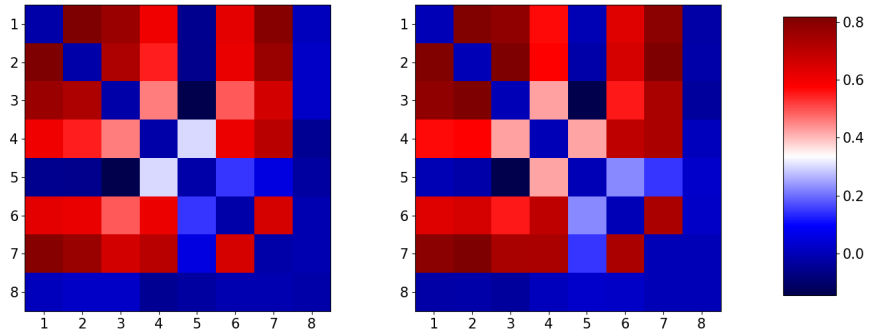
Figure 3: Comparison of the ground truth posterior of a linear regression model fit to the Boston Housing dataset, and the approximate posterior learned by VIFA (with Laplace prior). Variances and covariance are plotted separately due to the difference in their magnitude. In plot (c), the diagonal entries of the covariance matrices have been set to zero.



(a) Comparison of true and estimated posterior means

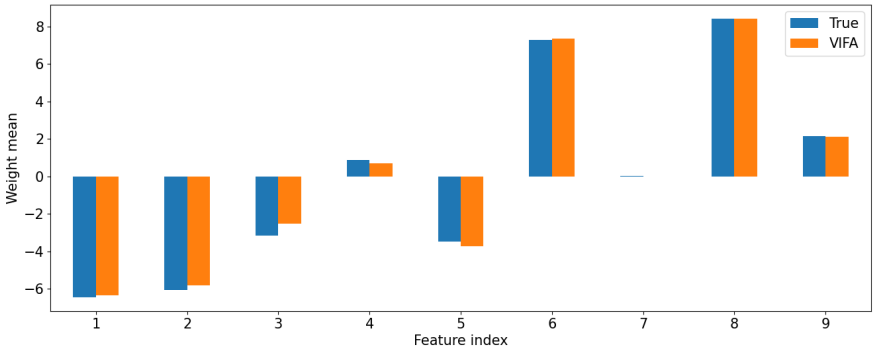


(b) Comparison of true and estimated posterior variances

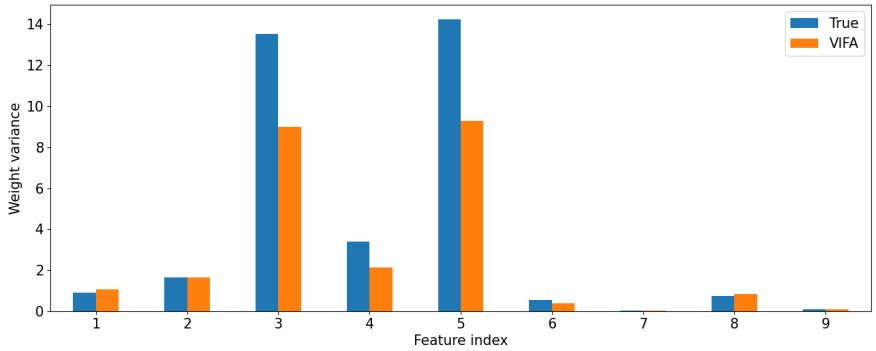


(c) Comparison of true and estimated posterior covariances

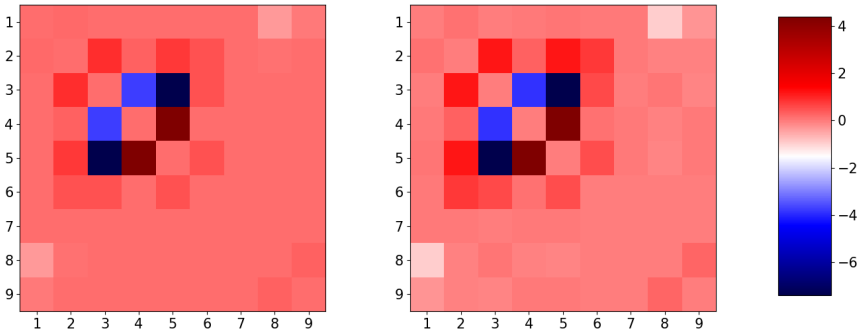
Figure 4: Comparison of the ground truth posterior of a linear regression model fit to the Concrete Strength dataset, and the approximate posterior learned by VIFA (with Laplace prior). Variances and covariance are plotted separately due to the difference in their magnitude. In plot (c), the diagonal entries of the covariance matrices have been set to zero.



(a) Comparison of true and estimated posterior means



(b) Comparison of true and estimated posterior variances



(c) Comparison of true and estimated posterior covariances

Figure 5: Comparison of the ground truth posterior of a linear regression model fit to the Concrete Strength dataset, and the approximate posterior learned by VIFA (with Laplace prior). Variances and covariance are plotted separately due to the difference in their magnitude. In plot (c), the diagonal entries of the covariance matrices have been set to zero.

on energy dataset, but it mildly fall behind Gaussian prior on other three datasets.

Metric	Dataset	SmallNN-VIFA-Gaussian	SmallNN-VIFA-Laplace
NMLL	Energy	2.53± 0.02	**2.51±0.03**
	Boston	**2.66±0.06**	2.68±0.07
	Concrete	**3.34±0.01**	3.35±0.01
	Yacht	**2.36±0.06**	2.39±0.08
RMSE	Energy	2.90±0.06	**2.83±0.06**
	Boston	**3.64±0.23**	3.75±0.26
	Concrete	**6.77±0.09**	6.87±0.10
	Yacht	**2.51±0.09**	2.58±0.08

Table 2: Comparision between VIFA with gaussian prior and VIFA with laplace prior applied on small neural network. The best results among all SmallNN experiments are highlighted in bold.

3.2 Probabilistic PCA

We test the performance of VIFA algorithm for larger neural networks (2 hidden layers, each with 50 neurons) with experimental results summarized in the second (LargeNN-VIFA-Gaussian) and the forth column (LargeNN-VIFA-Laplace) of the table 3 with gaussian and laplace priors respectively. By comparing these two columns to the results in table 2, we realize that larger neural network performs worse than smaller ones with VIFA algorithm, which reveals its overfitting issue.

To allivate this issue, we propose VIPPCA algorithm. We test its effectiveness for both Gaussian and Laplace prior scenarios. For gaussian prior cases, precision is log-uniformaly sampled from [0.01, 10] and for laplace prior cases, α is log-uniformaly sampled from [0.141, 4.472], which ensures the same range of prior precision for gaussian and laplace prior. Different from VIFA, we have a new hyperparameter σ, and we tune it by random search sampled log-uniformly from [0.01, 1].

The experiment results for VIPPCA algorithm is summarized in the first (LargeNN-VIPPCA-Gaussian) and the third (LargeNN-VIPPCA-Laplace) column in table 3. Comparison to the results of VIFA shows that VIPPCA can effective improves the performance when larger neural networks are used.

Metric	Dataset	LargeNN-VIPPCA-Gaussian	LargeNN-VIFA-Gaussian	LargeNN-VIPPCA-Laplace	LargeNN-VIFA-Laplace	SmallNN-VIFA-Gaussian	SmallNN-VIFA-Laplace
NMLL	Energy	2.43±0.03	3.03±0.08	**2.40±0.03**	3.29±0.07	2.53±0.02	**2.51±0.03**
	Boston	2.70±0.07	3.43±0.04	**2.63±0.04**	3.54±0.03	**2.66+0.06**	2.68+0.07
	Concrete	3.32±0.02	3.77±0.14	**3.29±0.02**	4.06±0.11	**3.34±0.01**	3.35±0.01
	Yacht	**2.34±0.03**	4.01±0.04	**2.34±0.03**	4.00±0.04	**2.36±0.06**	2.39±0.08
RMSE	Energy	2.66±0.11	4.20±0.45	**2.57±0.11**	5.55±0.63	2.90±0.06	**2.83±0.06**
	Boston	3.58±0.23	7.36±0.38	**3.55±0.20**	8.53±0.35	**3.64±0.23**	3.75±0.27
	Concrete	6.13±0.10	10.37±1.39	**6.10±0.14**	16.92±3.54	**6.77±0.09**	6.87±0.10
	Yacht	2.73±0.30	13.86±0.43	**2.61±0.27**	14.00±0.48	**2.51±0.09**	2.58±0.08

Table 3: Comparison of applying Probabilistic PCA variational distribution, and the factor analysis variational distribution on larger neural networks. VIFA refers to using factor analysis as variational distribution and VIPPCA stands for using probabilistic PCA as variational distribution. The best results among all LargeNN experiments are highlighted in bold in the left panel. The copy of table 2 is in the right panel.

A Experiment Results for Laplace prior $\in$ [0.01, 10]

For neural network experiments, if we do not restrict the range of the random search of precision of laplace prior same as the range for gaussian prior, we can set laplace prior range to be [0.01, 10] [2]

Metric	Dataset	VIFA-NN-Gaussian	VIFA-NN-Laplace
NMLL	Energy	2.53± 0.02	**2.20±0.03**
	Boston	**2.66±0.06**	2.67±0.08
	Concrete	**3.34±0.01**	3.36±0.01
	Yacht	**2.36±0.06**	**2.36±0.06**
RMSE	Energy	2.90±0.06	**2.10±0.05**
	Boston	**3.64±0.23**	3.67±0.27
	Concrete	**6.77±0.09**	6.89±0.09
	Yacht	**2.51±0.09**	2.70±0.11

Metric	Dataset	LargeNN-fixD-Gaussian	LargeNN-full-Gaussian	LargeNN-fixD-Laplace	LargeNN-full-Laplace	SmallNN-full-Gaussian	Small-full-Laplace
NMLL	Energy	2.43±0.03	3.03±0.08	2.37±0.03	2.91±0.06	2.53±0.02	2.20±0.03
	Boston	2.70±0.07	3.43±0.04	2.70±0.06	3.29±0.05	2.66+0.06	2.67+0.08
	Concrete	3.32±0.02	3.77±0.14	3.31±0.02	3.52±0.09	3.34±0.01	3.36±0.01
	Yacht	2.34±0.03	4.01±0.04	2.32±0.03	4.00±0.04	2.36±0.06	2.36±0.06
RMSE	Energy	2.66±0.11	4.20±0.45	2.49±0.12	4.14±0.48	2.90±0.06	2.10±0.05
	Boston	3.58±0.23	7.36±0.38	3.54±0.25	6.24±0.52	3.64±0.23	3.67±0.27
	Concrete	6.13±0.10	10.37±1.39	6.14±0.13	8.73±1.27	6.77±0.09	6.89±0.09
	Yacht	2.73±0.30	13.86±0.43	2.46±0.28	13.68±0.42	2.51±0.09	2.70±0.11

[2] this suggest the actual precision range laplace prior uses is $[5 \times 10^{-5}, 50]$, larger than the range used by gaussian prior, which is unfair for gaussian prior.

References

[1] Mohsen Joneidi. Sparse auto-regressive: Robust estimation of ar parameters. 06 2013.