doi:10.20944/preprints202208.0234.v1

Article

Variational Bayesian Approximation (VBA): A comparison between three optimization algorithms

Azadeh Fallah[1] and Ali Mohammad-Djafari [2,3,*] orcid number:0000-0003-0678-7759

[1] Ferdowsi University of Mashhad, Mashhad, Iran; azadeh.fallah@mail.um.ac.ir
[2] International Science Consulting and Training (ISCT), 91440 Bures sur Yvette, France; djafari@ieee.org
[3] Scientific leader of Shanfeng company, Shaoxing, China
Correspondence: azadeh.fallah@mail.um.ac.ir

Abstract: In many Bayesian computations, first, we obtain the expression of the joint distribution of all the unknown variables given the observed data. In general, this expression is not separable in those variables. Thus, obtaining their marginals for each variable and computing the expectations are difficult and costly. This problem becomes even more difficult in high dimensional quandaries, which is an important issue in inverse problems. We may then try to propose a surrogate expression with which we can do approximate computations. Often a separable expression approximation can be useful enough. The Variational Bayesian Approximation (VBA) is a technique that approximates the joint distribution p with an easier, for example separable, one q by minimizing Kullback–Leibler Divergence $KL(q|p)$. When q is separable in all the variables, the approximation is also called Mean Field Approximation (MFA) and so q is the product of the approximated marginals. A first standard and general algorithm is alternate optimization of $KL(q|p)$ with respect to q_i. A second general approach is its optimization in the Riemannian manifold. However, in this paper, for practical reasons, we consider the case where p is in the exponential family and so is q. For this case, $KL(q|p)$ becomes a function of the parameters $\boldsymbol{\theta}$ of the exponential family. Then, we can use any other optimization algorithm to obtain those parameters. In this paper, we compare three optimization algorithms: standard alternate optimization, a gradient-based algorithm and a natural gradient algorithm and study their relative performances on three examples.

Keywords: Variational Bayesian Approach (VBA); Kullback–Leibler Divergence; Mean Field Approximation (MFA); Optimization Algorithm

1. Introduction

In many applications, with direct or undirect observations, the use of the Bayesian computations starts with obtaining the expression of the joint distribution of all the unknown variables given the observed data. Then, we have to use it to do inference. In general, this expression is not separable in all the variables of the problem. So, the computations becomes hard and costly. For example, obtaining the marginals for each variable and computing the expectations are difficult and costly. This problem becomes even more crutial in high dimensional quandaries, which is an important issue in inverse problems. We may then need to propose a surrogate expression with which we can do approximate computations.

The Variational Bayesian Approximation (VBA) is a technique that approximates the joint distribution p with an easier, for example a separable one q by minimizing Kullback–Leibler Divergence $KL(q|p)$, which makes the marginal computations much easier. For example, in the case of two variables, $p(x,y)$ is approximated by $q(x,y) = q_1(x)q_2(y)$ via minimizing $KL(q_1q_2|p)$. When q is separable in all the variables of p, the approximation is also called Mean Field Approximation (MFA).

To obtain the approximate marginals q_1 and q_2 we have to minimize $KL(q_1q_2|p)$. A first standard and general algorithm is alternate optimization of $KL(q_1q_2|p)$ with respect to q_1 and q_2. Finding the expression of the functional derivatives of $KL(q_1q_2|p)$ with respect to q_1 and q_2 and equating them to zero alternatively, we obtain an iterative optimization algorithm. A second general approach is its optimization in the Riemannian manifold. However, in this paper, for practical reasons, we consider the case where p is in the exponential family and so are q_1 and q_2. For this case, $KL(q_1q_2|p)$ becomes a function of the parameters $\boldsymbol{\theta}$ of the exponential family. Then, we can use any other optimization algorithm to obtain those parameters.
In this paper, we compare three optimization algorithms: standard alternate optimization (Algorithm 1), a gradient-based algorithm [1,2] (Algorithm 2) and a natural gradient algorithm [3–5] (Algorithm 3).

Between the main advantages of the VBA for inference problems, such as inverse problems and machine learning, we can mention the following:
- First, VBA builds a sufficient model according to prior information and the final posterior distribution. Especially in the Mean-Field Approximation (MFA), the result ends in an explicit form for each unknown component using conjugate priors, and works well for small sample sizes, [6–8].
- The second benefit is, for example in machine learning, a robust way for classification based on the predictive posterior distribution and diminishes parameter over-trained, [7].
- The third privilege is that the target structure has uncertainty in the VBA recursive processes. This feature prevents further error propagation and increases the robustness of VBA, [9].

Besides all these preponderance, the VBA has some weaknesses, such as difficulty in the solution of integrals and expectations to get a posterior distribution, and there is no evidence to find an exact posterior, [6]. Its most significant drawback arises when there are strong dependencies between unknown parameters, and the VBA ignores them. Then estimates, computed based on this approximation, may be very far from the exact values. However, it works well when the amounts of dependencies are low [8].
In this article, we work on three different estimating algorithms of the unknown parameters in a model concerning prior information. The first iterative algorithm is standard alternate optimization based on VBA, which start from some initial points. Sometimes, the points are estimated from an available dataset, but most of the time, we do not have enough data on the parameters to make some pre-estimation of them. To solve this obstacle, we can start the algorithm with some desired points, and then by repeating the process, they approach the true values using the posterior distribution. The second two algorithms are gradient-based and natural gradient algorithms, whose based function are Kullback–Leibler divergence. First, the gradient of Kullback–Leibler for all unknown parameters have to be found, then start from some points either estimated from data or desired choices. Then, we repeat the iterative algorithm till it convergences to some points. If we denote the unknown parameter space with $\boldsymbol{\theta}$, then the recursive formula is $\tilde{\boldsymbol{\theta}}^{(k+1)} = \tilde{\boldsymbol{\theta}}^{(k)} - \Delta KL(\tilde{\boldsymbol{\theta}}^{(k)})$ for gradient-based algorithm. The formula for natural gradient is pretty similar and is $\tilde{\boldsymbol{\theta}}^{(k+1)} = \tilde{\boldsymbol{\theta}}^{(k)} - \frac{\Delta KL(\tilde{\boldsymbol{\theta}}^{(k)})}{\|\Delta KL(\tilde{\boldsymbol{\theta}}^{(k)})\|}$.

Also, we consider three examples, Normal-Inverse-Gamma, multivariate Normal, and linear inverse problem for checking the performance and convergence speed of the algorithms.

We propose the following organization of this paper: In section 2, we present a brief explanation of the basic VBA analytical part. In section 3, we explain our first example related to Normal-Inverse-Gamma distribution analytically and, in practice, explain the outcomes of three algorithms to estimate the unknown parameters. In section 4, we study a more complex example of a multivariate Normal distribution whose means and variance-covariance matrix are unknown and have Normal Inverse Wishart distribution. The aim of this section is to demonstrate marginal distributions of $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ using a set of multivariate Normal observations using these mean and variance. In section 5, the example is more close to realistic situations and is a linear inverse problem. In section 6, we present a summary of the work done in the article and compare the three recursive algorithms in three different examples.

2. Variational Bayesian Approach (VBA)

As we mentioned earlier, VBA uses Kullback–Leibler Divergence. Kullback–Leibler Divergence [10] $KL(q|p)$ is an information measure of discrepancy between two probability functions defined as follow. Let $p(x)$ and $q(x)$ be two density functions of a continuous random variable x respect to support set $\mathbb{S}_X$. $KL(q|p)$ function is introduced as

$$KL(q|p) = \int_{x \in \mathbb{S}_X} q(x) \ln \frac{q(x)}{p(x)} \, \mathrm{d}x. \tag{1}$$

For simplicity, we assume a bivariate case of distribution $p(x, y)$, and want to assess it via VBA, then we have:

$$KL(q|p) = -H(q_1) - H(q_2) - < \ln\ p(x,y) >_{q_1 q_2}, \tag{2}$$

where

$$H(q_1) = -\int_{x \in \mathbb{S}_X} q_1(x) \ln\ q_1(x) \, \mathrm{d}x \quad \text{and} \quad H(q_2) = -\int_{y \in \mathbb{S}_Y} q_2(y) \ln\ q_2(y) \, \mathrm{d}y$$

are, respectively, the Shannon entropies of x and of y, and

$$< \ln p(x,y) >_{q_1 q_2} = \int \int_{(x,y) \in \mathbb{S}_{XY}} q_1(x) q_2(y) \ln\ p(x,y) \, \mathrm{d}x \, \mathrm{d}y.$$

Now, differentiating the equation (2) with respect to q_1, and then with respect to q_2 and equating them to zero, we obtain:

$$q_1(x) \propto \exp\left\{ < \ln\ p(x,y) >_{q_2(y)} \right\} \quad \text{and} \quad q_2(y) \propto \exp\left\{ < \ln\ p(x,y) >_{q_1(x)} \right\} \tag{3}$$

These results can be easily extended to more dimensions [11]. They do not have any closed form, because they depend on the expression of $p(x, y)$ and those of q_1 and q_2. An interesting case is the case of exponential families and conjugate priors, where writing

$$p(x,y) = p(x|y)p(y), \text{and } p(y|x) = \frac{p(x,y)}{p(x)} = \frac{p(x|y)p(y)}{p(x)}, \tag{4}$$

we can consider $p(y)$ as prior, $p(x|y)$ as the likelihood, and $p(y|x)$ as the posterior distributions. Then, if $p(y)$ is a conjugate prior for the likelihood $p(x|y)$, then the posterior $p(y|x)$ will be in the same family as the prior $p(y)$. To illustrate all these properties, we give details of these expressions for a first simple example of Normal-Inverse-Gamma $p(x, y) = \mathcal{N}(x|\mu, y)\mathcal{IG}(y|\alpha, \beta)$ with $q_1(x) = \mathcal{N}(x|\mu, v)$ and $q_2 = \mathcal{IG}(y|\alpha, \beta)$. For this simple case, first we give the expression of $KL(q|p)$ with $q_1(x) = \mathcal{N}(x|\widetilde{\mu}, \widetilde{v})$ and $q_2(y) = \mathcal{IG}(y|\widetilde{\alpha}, \widetilde{\beta})$ as a function of the parameters $\boldsymbol{\theta} = (\widetilde{\mu}, \widetilde{v}, \widetilde{\alpha}, \widetilde{\beta})$ and then the expressions of the three above-mentioned algorithms and we study their convergence.

3. Normal-Inverse-Gamma Distribution Example

The purpose of this section is to explain in detail the process of performing calculations in VBA. For this we consider a simple case for which, we have all the necessary expressions. The objective here is to compare the three different algorithms mentioned earlier. Also, its practical application can be explained as follows:

We have a sensor which measures a quantity X, N times $x_1, ..., x_N$. We want to model these data. In a first step, we model it as $N(x|\mu, v)$ with fixed μ and v. Then, it is easy to estimate the parameters (μ, v) either by Maximal Likelihood or Bayesian strategy. If we assume that the model is Gaussian with unknown variance and call this variance y and assign an $\mathcal{IG}$ prior to it, then we have a model

$\mathcal{NIG}$ for $p(x,y)$. We showed that the margins are St and $\mathcal{IG}$. Working directly with St is difficult. So, we want to approximate it with a Gaussian $q_1(x)$. This is equivalent to approximating $p(x,y)$ with $q_1(x)q_2(y)$. Now, we want to find the parameters μ, v, α, and β, which minimize $KL(q_1q_2|p)$. This process is called VBA. Then, we want to compare three algorithms to obtain the parameters which minimize $KL(\cdot|\cdot)$. $KL(\cdot|\cdot)$ is convex with respect to q_1 if q_2 is fixed and is convex with respect to q_2 if q_1 is fixed. So, we hope that the iterative algorithm converges. However, $KL(\cdot|\cdot)$ may not be convex in the space of parameters. So, we have to study the shape of this criterion concerning the parameters $\widetilde{v}$, $\widetilde{\alpha}$, and $\widetilde{\beta}$.

The practical problem considered here is in the following: A sensors delivers a few samples $\boldsymbol{x} = \{x_1, x_2, \cdots, x_N\}$ of a physical quantity X. We want to find $p(x)$. For this process, we assume a simple Gaussian model, but with unknown variance y. So that, the forward model can be written as $p(x,y) = \mathcal{N}(x|\mu,y)\mathcal{IG}(y|\alpha,\beta)$. In this simple example, we know that $p(x)$ is a Student-t distribution obtained by:

$$S(x|m,\alpha,\beta) = \int \mathcal{N}(x|\mu,y)\mathcal{IG}(y|\alpha,\beta)\,\mathrm{d}y \tag{5}$$

Our objective is to find the three parameters $\boldsymbol{\theta} = (\mu,\alpha,\beta)$ from the data $\boldsymbol{x}$ and an approximate marginal $q(x)$ for $p(x)$.

The main idea is to find such $q_1(x)q_2(y)$ as an approximation of $p(x,y)$. Here, we show the VBA, step by step. For this, we start by choosing the conjugate families $q_1(x) = \mathcal{N}(x|\widetilde{\mu},\widetilde{v})$ and $q_2(y) = \mathcal{IG}(y|\widetilde{\alpha},\widetilde{\beta})$.

In the first step, we have to calculate $\ln p(x,y)$

$$\ln p(x,y) = c - \frac{1}{2}\ln y - \frac{1}{2y}(x-\widetilde{\mu})^2 - (\widetilde{\alpha}+\frac{1}{2})\ln y - \frac{\widetilde{\beta}}{y}. \tag{6}$$

where c is a constant value term independent of x and y. First of all, we have to start by finding q_2, so the integration of $\ln p(x,y)$ is concerning q_1

$$< \ln\ p(x,y) >_{q_1} = c - \frac{1}{2y} < (x-\widetilde{\mu})^2 >_{q_1} -(\widetilde{\alpha}+1)\ln\ y - \frac{\widetilde{\beta}}{y}. \tag{7}$$

Since the mean of x is the same in prior and posterior distribution, $\widetilde{\mu} = \widetilde{\mu}'$, and then $< (x-\widetilde{\mu})^2 >_{q_1} = \widetilde{v}$. Thus

$$q_2(y) \propto \exp\left[-(\widetilde{\alpha}+1)\ln\ y - (\frac{\widetilde{v}}{2}+\widetilde{\beta})\frac{1}{y}\right]. \tag{8}$$

Thus, the function $q_2(y)$ is equivalent to an inverse gamma distribution $\mathcal{IG}(\widetilde{\alpha}, \frac{\widetilde{v}}{2}+\widetilde{\beta})$. Now, we have to take integral of $\ln\ p(x,y)$ over q_2 to find q_1

$$< \ln\ p(x,y) >_{q_2} = c - (\widetilde{\alpha}+1) < \ln\ y >_{q_2} -(\widetilde{\beta}+\frac{1}{2}(x-\widetilde{\mu})^2) < \frac{1}{y} >_{q_2} \tag{9}$$

Note that the first term does not depend on x and in the second term we have $< \frac{1}{y} >_{q_2} = \frac{2\widetilde{\alpha}}{2\widetilde{\beta}+\widetilde{v}}$, so

$$q_1(x) \propto \exp\left[-\frac{2\widetilde{\alpha}}{2\widetilde{\beta}+\widetilde{v}}(\widetilde{\beta}+\frac{1}{2}(x-\widetilde{\mu})^2)\right] \propto \exp\left[-\frac{(x-\widetilde{\mu})^2}{2\frac{2\widetilde{\beta}+\widetilde{v}}{2\widetilde{\alpha}}}\right]. \tag{10}$$

We see that q_1 is again a Normal distribution but with updated parameters $\mathcal{N}(\widetilde{\mu}, \frac{2\widetilde{\beta}+\widetilde{v}}{2\widetilde{\alpha}})$, so $\widetilde{v} = \frac{2\widetilde{\beta}+\widetilde{v}}{2\widetilde{\alpha}}$. Note that, we obtained the conjugacy property: If $p(x|y) = \mathcal{N}(x|\mu,y)$ and $p(y) = \mathcal{IG}(y|\alpha,\beta)$, then $p(y|x) = \mathcal{IG}(y|\alpha',\beta')$ where μ', α' and β' are $\mu' = \mu$, $\alpha' = \alpha$, $\beta' = \beta + \frac{2\beta+v}{4\alpha}$. In this case, we also know that $p(x|\alpha,\beta) = St(x|\mu',\alpha,\beta)$.

In standard alternate optimization (algorithm 1), there is no need for an iterative process for $\widetilde{\mu}$ and $\widetilde{\alpha}$, which are approximated by $\widetilde{\mu} = \mu_0$ and $\widetilde{\alpha} = \alpha_0$, respectively. The situation for $\widetilde{\beta}$ and $\widetilde{v}$ is different because there are circular dependencies among them. So, the approximation needs an iterative process, staring from $\widetilde{\mu}^{(1)} = \mu_0$, $\widetilde{v}^{(1)} = v_0$, $\alpha^{(1)} = \alpha_0$, and $\beta^{(1)} = \beta_0$. The main conclusion for this case is that the VBA algorithm becomes:

Algorithm 1:

$$\begin{aligned}
\widetilde{\alpha}^{(k+1)} &= \widetilde{\alpha}^{(k)}, \\
\widetilde{\beta}^{(k+1)} &= \widetilde{\beta}^{(k)} + \frac{\widetilde{v}^{(k)}}{2}, \\
\widetilde{\mu}^{(k+1)} &= \widetilde{\mu}^{(k)}, \\
\widetilde{v}^{(k+1)} &= \frac{2\widetilde{\beta}^{(k)} + \widetilde{v}^{(k)}}{2\widetilde{\alpha}^{(k)}}.
\end{aligned}$$

For the two other algorithms, gradient and natural gradient-based, it requires to find the expression of $KL(q_1q_2 : p)$ as a function of the parameters $\boldsymbol{\theta} = (\alpha, \beta, \mu, v)$

$$KL(\tilde{\boldsymbol{\theta}}) = 2\ln\,\Gamma(\tilde{\alpha} - \frac{1}{2}) - (2\widetilde{\alpha} + \frac{3}{2})\psi_0(\widetilde{\alpha} - \frac{1}{2}) + \frac{(2\widetilde{\alpha} - 1)(\widetilde{v} + 2\widetilde{\beta})}{4\widetilde{\beta}} + \widetilde{\alpha} + \frac{5}{2}\ln\widetilde{\beta} - \frac{1}{2}\ln\,\tilde{v} - 1, \quad (11)$$

Then, we also need the gradient expression of $\nabla KL(\widetilde{\boldsymbol{\theta}})$ for $\widetilde{\boldsymbol{\theta}}$

$$\nabla KL(\tilde{\boldsymbol{\theta}}) = \left(\frac{\widetilde{v}}{2\widetilde{\beta}} - (2\widetilde{\alpha} + \frac{3}{2})\psi_1(\widetilde{\alpha} - \frac{1}{2}) + 2, \quad -(2\widetilde{\alpha} - 1)\frac{\widetilde{v}}{4\widetilde{\beta}^2} + \frac{5}{2\widetilde{\beta}}, \quad 0, \quad \frac{2\widetilde{\alpha} - 1}{4\widetilde{\beta}} - \frac{1}{2\widetilde{v}} \right), \quad (12)$$

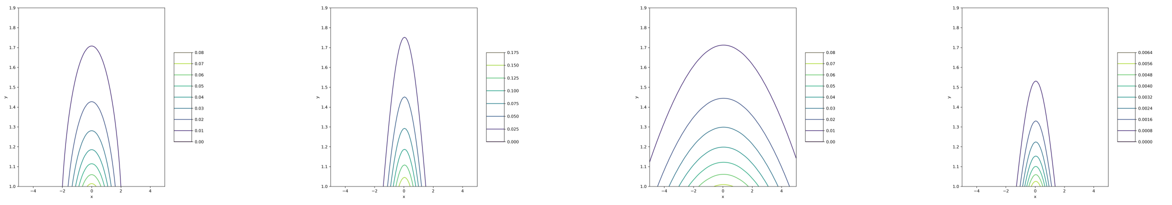
As we can see these expressions do not depend on $\widetilde{\mu}$, so their derivatives are zero. The means are preserved.

Algorithm 2 and 3:

$$\begin{aligned}
\widetilde{\alpha}^{(k+1)} &= \widetilde{\alpha}^{(k)} - \gamma \left(\frac{\widetilde{v}^{(k)}}{2\widetilde{\beta}^{(k)}} - (2\widetilde{\alpha}^{(k)} + \frac{3}{2})\psi_1(\widetilde{\alpha}^{(k)} - \frac{1}{2}) + 2 \right), \\
\widetilde{\beta}^{(k+1)} &= \widetilde{\beta}^{(k)} - \gamma \left(-(2\widetilde{\alpha}^{(k)} - 1)\frac{\widetilde{v}^{(k)}}{4[\widetilde{\beta}^{(k)}]^2} + \frac{5}{2\widetilde{\beta}^{(k)}} \right), \\
\widetilde{\mu}^{(k+1)} &= \widetilde{\mu}^{(k)}, \\
\widetilde{v}^{(k+1)} &= \widetilde{v}^{(k+1)} - \gamma \left(\frac{2\widetilde{\alpha}^{(k)} - 1}{4\widetilde{\beta}^{(k)}} - \frac{1}{2\widetilde{v}^{(k)}} \right),
\end{aligned}$$

where γ is fixed for the gradient algorithm, and is proportional to $1/\|\nabla KL\|$ for the natural gradient algorithm. To start the numerical computations, we generate $n = 100$ samples from the model $p(x, y) = \mathcal{N}(x|1, y)\mathcal{IG}(y|3, 1)$. Thus, we know the exact values of the unknown parameters, just keeping in mind, not used in algorithms. $\tilde{\boldsymbol{\theta}}_1$, $\tilde{\boldsymbol{\theta}}_1$, and $\tilde{\boldsymbol{\theta}}_1$ are the estimated parameters using the alternative, gradient-based, and natural gradient algorithms along with their surface and contour plots in fig 1, respectively

$$\tilde{\boldsymbol{\theta}}_1 = \begin{cases} \widetilde{\mu} = 0.044724, \\ \widetilde{v} = 0.590894, \\ \widetilde{\alpha} = 3.792337, \\ \widetilde{\beta} = 1.765735 \end{cases} \quad \tilde{\boldsymbol{\theta}}_2 = \begin{cases} \widetilde{\mu} = 0.044724, \\ \widetilde{v} = 7.910504, \\ \widetilde{\alpha} = 4.991621, \\ \widetilde{\beta} = 3.002594 \end{cases} \quad \tilde{\boldsymbol{\theta}}_3 = \begin{cases} \widetilde{\mu} = 0.044724, \\ \widetilde{v} = 0.423761, \\ \widetilde{\alpha} = 4.415239, \\ \widetilde{\beta} = 0.706049. \end{cases}$$



(a) True contour plot of the model

(b) The VBA approximation

(c) The gradient-based algorithm

(d) The natural gradient algorithm

Figure 1. The true model has $\mathcal{N}(x|0,y)\mathcal{IG}(y|3,1)$. The numbers of convergence are diverse in the algorithms.

All three algorithms try to minimize the same criterion. So, the objectives are all the same, but the number of steps may differ. The requirements must reach the minimum $KL(\cdot)$. In this simple example of the Normal-Inverse-Gamma distribution, the convergence step numbers of VBA, gradient-based, and natural gradient are 1, 2, and 1. The overall performance of the standard alternate optimization (VBA) is more precise than any others. The poorest estimation is from a gradient-based algorithm. So, the algorithms are able to approximate the joint density function with a separable one, but with different accuracy. In the following section, we will tackle a more complex model.

4. Multivariate Normal Inverse Wishart Example

In previous section, we explain how to preform VBA to approximate a complicated joint distribution function by tractable marginal factorials over a simple case study. In this section, a multivariate Normal case $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$ is considered which is approximated by $q(\boldsymbol{x}) = \prod_i \mathcal{N}(x_i|\widetilde{\mu}_i, \widetilde{v}_i)$ for different shapes for the covariance matrix $\widetilde{\boldsymbol{\Sigma}}$.
We assume that the basic structure of an available data set is multivariate Normal with unknown mean vector $\widetilde{\boldsymbol{\mu}}$ and variance-covariance matrix $\widetilde{\boldsymbol{\Sigma}}$. Their joint prior distribution is a Normal Inverse Wishart distribution of $\mathcal{NIW}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}}|\widetilde{\boldsymbol{\mu}}_0, \widetilde{\kappa}, \widetilde{\boldsymbol{\Psi}}, \widetilde{v})$. The posteriors are multivariate Normal for mean vector and Inverse Wishart for variance-covariance matrix. Since the Normal Inverse Wishart distribution is a conjugate prior distribution for multivariate Normal, the posterior distribution of $\widetilde{\boldsymbol{\mu}}$ and $\widetilde{\boldsymbol{\Sigma}}$ is again belong to the same family, and their corresponding margins are

$$\mathcal{MN}\left(\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0 + n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n}, \frac{1}{\tilde{\kappa}+n}\tilde{\boldsymbol{\Lambda}}\right), \quad \mathcal{IW}\left(\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}} + \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T, \tilde{v}+n\right), \tag{13}$$

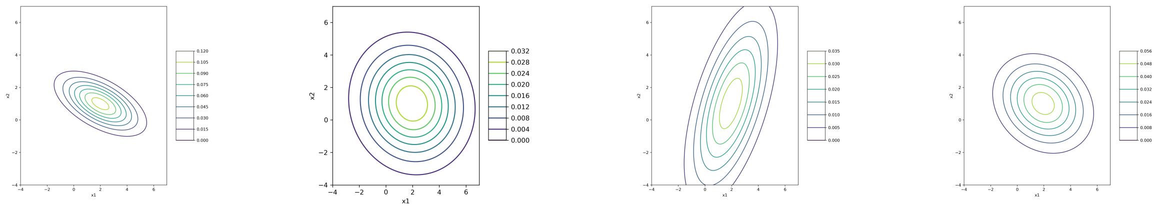
where n is the sample size.To present the performance of the three algorithms, we work on a data set coming from $\boldsymbol{x} \sim \mathcal{NIW}(\boldsymbol{x}|\boldsymbol{\mu}, \,\Sigma)$ whose parameters have the below density structure

$$\boldsymbol{\mu} \sim \mathcal{MN}\left(\boldsymbol{\mu}|\begin{bmatrix}2\\1\end{bmatrix}, \frac{1}{2}\begin{bmatrix}3 & -1\\-1 & 1\end{bmatrix}\right), \quad \Sigma \sim \mathcal{IW}\left(\Sigma|\begin{bmatrix}3 & -1\\-1 & 1\end{bmatrix}, 6\right).$$

We use only the data of $\boldsymbol{x}$ in the estimation processes. The results of algorithms are drowned in fig 2 along with the true contour plot of the model. The VBA estimation is the most separable distribution compared with gradient and natural gradient methods. The next best case is the natural gradient algorithm, but its weakness is transferring the dependency a little bit to the approximation. The result for the gradient-based is completely shown the dependency, and its inability to get a separable model.

5. Simple Linear Inverse Problem

Finally the third example is the case of linear inverse problems with $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \boldsymbol{\epsilon}$ with priors $p(\boldsymbol{\epsilon}) = \mathcal{N}(\boldsymbol{\epsilon}|0, v_\epsilon \boldsymbol{I})$ and $p(\boldsymbol{f}) = \mathcal{N}(\boldsymbol{f}|0, \text{diag}\,[\boldsymbol{v}])$ with $\boldsymbol{f} = [f_1, f_2, \cdots, f_N]$ and $\boldsymbol{v} = [v_{f_1}, v_{f_2}, \cdots, v_{f_N}]$



(a) True contour plot of the model

(b) The VBA approximation

(c) The gradient-based algorithm

(d) The natural gradient algorithm

Figure 2. The number of steps to get the result for 2b, 2c, and 2d are 2, 280, and 4, respectively. In this example, the convergence numbers are different between three algorithms, in the gradient-based. Although the gradient-based is not able to approximate the joint density function with a separable one, the VBA and natural gradient algorithms estimate the distribution with separable ones in fewer steps.

for which we have $p(\boldsymbol{f},\boldsymbol{v}|\boldsymbol{g}) \propto p(\boldsymbol{g}|\boldsymbol{f},v_\epsilon)p(\boldsymbol{f}|\boldsymbol{v})p(\boldsymbol{v})$ with $p(\boldsymbol{g}|\boldsymbol{f},v_\epsilon) = \mathcal{N}(\boldsymbol{g}|\boldsymbol{H}\boldsymbol{f},v_\epsilon\boldsymbol{I})$, $p(\boldsymbol{f}|\boldsymbol{v}) = \mathcal{N}(\boldsymbol{f}|0,\text{diag}\,[\boldsymbol{v}])$ and $p(\boldsymbol{v}|\alpha,\beta) = \prod_j \mathcal{IG}(v_j|\alpha,\beta)$, [12]. Thus, the joint distribution of $\boldsymbol{g}$, $\boldsymbol{f}$, and $\boldsymbol{v}$ is estimated by VBA as the following relation

$$p(\boldsymbol{g},\boldsymbol{f},\boldsymbol{v}) \propto p(\boldsymbol{g}|\boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{v})p(\boldsymbol{v}). \tag{14}$$

Although we can mathematically compute the margins in this particular example, we desire to approximate them via the iterative VBA algorithm $q(\boldsymbol{g},\widetilde{\boldsymbol{f}},\tilde{\boldsymbol{v}}) = q_1(\boldsymbol{g}|\widetilde{\boldsymbol{f}})q_2(\widetilde{\boldsymbol{f}})q_3(\tilde{\boldsymbol{v}})$ compared them with gradient and natural gradient-based algorithms. The objective function is the estimation of $q_2(\widetilde{\boldsymbol{f}})$, but in the recursive process, $q_1(\boldsymbol{g}|\widetilde{\boldsymbol{f}})$ and $q_3(\tilde{\boldsymbol{v}})$ are updated, too. For simplicity, we suppose that the transposition matrix $\boldsymbol{H}$ is an identical matrix $\boldsymbol{I}$. The final outputs are as follows

$$\begin{gathered}\widetilde{\boldsymbol{f}} \sim \mathcal{MN}\left(\frac{\widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}}{1+\frac{2\tilde{v}_\epsilon\tilde{\boldsymbol{\alpha}}}{n(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\widetilde{f}}+2\tilde{\boldsymbol{\beta}})}}, \text{diag}\left[\frac{\tilde{v}_\epsilon(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\widetilde{f}}+2\tilde{\boldsymbol{\beta}})}{n(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\widetilde{f}})+2n\tilde{\boldsymbol{\beta}}+2\tilde{v}_\epsilon\tilde{\boldsymbol{\alpha}}}\right]\right), \\ \boldsymbol{g} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}},\tilde{v}_\epsilon\boldsymbol{I}), \quad \text{and} \quad \tilde{v}_k \sim \mathcal{IG}\left(\tilde{\alpha}_k, \frac{\tilde{v}_k+\widetilde{\mu}^2_{\widetilde{f}_k}}{2}+\widetilde{\beta}_k\right), \; k = 1,\cdots,p.\end{gathered} \tag{15}$$

We choose a model to see the performance of these margins and compare them with gradient and natural gradient algorithms. The selected model is $\boldsymbol{g} = \boldsymbol{H}\boldsymbol{f} + \epsilon$ with the following knowledge

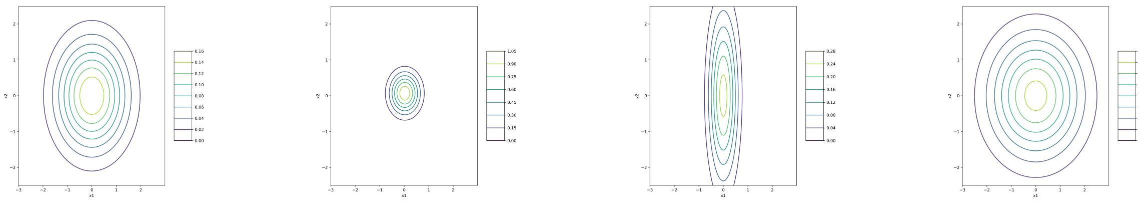
$$\boldsymbol{H} = \boldsymbol{I}, \quad \boldsymbol{f} \sim \mathcal{MN}(\boldsymbol{f}|\boldsymbol{0},\text{diag}\,[v_1,v_2]), \quad v_1 \sim \mathcal{IG}(v_1|3,2), \quad v_2 \sim \mathcal{IG}(v_2|4,3), \quad \epsilon \sim \mathcal{MN}(\epsilon|\boldsymbol{0},\boldsymbol{I}).$$

In the assessment procedure, we do not apply the above information. The output of algorithms are shown in fig 3, as well as the actual contour plot. In this example, the best diagnosis is from the natural gradient algorithm. The VBA by construction is separable and cannot be the same as the original.

6. Conclusion

This paper presents an approximation method of the unknown density functions for hidden variables called VBA. It is compared with gradient and natural gradient algorithms. We also consider three examples Normal Inverse Gamma, Normal Inverse Wishart, and linear inverse problem. We put the details of the first model here and give the details for two other examples in the whole paper. In all three models, the parameters are unexplored and need to be estimated by recursive algorithms. We try to approximate the joint complex distribution with a simpler version of the margin factorials that looks like independent cases. The VBA and natural gradient converge pretty soon. The major discrepancy in

doi:10.20944/preprints202208.0234.v1



(a) True contour plot of the model

(b) The VBA approximation

(c) The gradient-based algorithm

(d) The natural gradient algorithm

Figure 3. The true model is almost separable. The natural gradient algorithm works better in this example, and the poorest approximation is for the gradient-based algorithm.

algorithms comes from the accuracy of the results. They estimate the intricate joint distribution with separable ones. The overall performance of VBA is the best here.

References

1. Acerbi, L. Variational bayesian monte carlo. *Advances in Neural Information Processing Systems* **2018**, *31*.
2. Kuusela, M.; Raiko, T.; Honkela, A.; Karhunen, J. A gradient-based algorithm competitive with variational Bayesian EM for mixture of Gaussians. 2009 International Joint Conference on Neural Networks. IEEE, 2009, pp. 1688–1695.
3. Gharsalli, L.; Duchêne, B.; Mohammad-Djafari, A.; Ayasso, H. A gradient-like variational Bayesian approach: Application to microwave imaging for breast tumor detection. 2014 IEEE International Conference on Image Processing (ICIP). IEEE, 2014, pp. 1708–1712.
4. Zhang, G.; Sun, S.; Duvenaud, D.; Grosse, R. Noisy natural gradient as variational inference. International Conference on Machine Learning. PMLR, 2018, pp. 5852–5861.
5. Lin, W.; Khan, M.E.; Schmidt, M. Fast and simple natural-gradient variational inference with mixture of exponential-family approximations. International Conference on Machine Learning. PMLR, 2019, pp. 3992–4002.
6. Ghahramani, Z.; Beal, M. Variational inference for Bayesian mixtures of factor analysers. *Advances in neural information processing systems* **1999**, *12*.
7. Watanabe, S.; Minami, Y.; Nakamura, A.; Ueda, N. Variational Bayesian estimation and clustering for speech recognition. *IEEE Transactions on Speech and Audio Processing* **2004**, *12*, 365–381.
8. Montesinos-López, O.A.; Montesinos-López, A.; Crossa, J.; Montesinos-López, J.C.; Luna-Vázquez, F.J.; Salinas-Ruiz, J.; Herrera-Morales, J.R.; Buenrostro-Mariscal, R. A variational Bayes genomic-enabled prediction model with genotype× environment interaction. *G3: Genes, Genomes, Genetics* **2017**, *7*, 1833–1853.
9. Babacan, S.D.; Molina, R.; Katsaggelos, A.K. Variational Bayesian super resolution. *IEEE Transactions on Image Processing* **2010**, *20*, 984–999.
10. Kullback, S.; Leibler, R.A. On information and sufficiency. *The annals of mathematical statistics* **1951**, *22*, 79–86.
11. Fox, C.W.; Roberts, S.J. A tutorial on variational Bayesian inference. *Artificial intelligence review* **2012**, *38*, 85–95.
12. Ayasso, H.; Mohammad-djafari, A. Joint image restoration and segmentation using Gauss-Markov-Potts prior models and variational Bayesian computation. 2009 16th IEEE International Conference on Image Processing (ICIP), 2009, pp. 1297–1300. doi:10.1109/ICIP.2009.5413589.
13. Gupta, M.; Srivastava, S. Parametric Bayesian estimation of differential entropy and relative entropy. *Entropy* **2010**, *12*, 818–843.

182 Appendix A Computation of $KL(\tilde{\boldsymbol{\theta}})$ and its Gradient

The mathematical computations of Kullback–Leibler divergence $KL(\tilde{\boldsymbol{\theta}})$ in section 3 is based on (2) as follows

$$\begin{aligned}KL(q_1q_2|p) =& - H(q_1) - H(q_2) - \int\int q_1(x)q_2(y) \ln \left\{ \frac{1}{\sqrt{2\pi y}} \exp\{-\frac{(x-\tilde{\mu})^2}{2y}\} \right. \\ & \left. \frac{\widetilde{\beta}^{\widetilde{\alpha}-\frac{1}{2}}}{\Gamma(\widetilde{\alpha}-\frac{1}{2})} y^{-(\widetilde{\alpha}+\frac{1}{2})} \exp\{-\frac{\widetilde{\beta}}{y}\} \right\} dxdy \\ =& - H(q_1) - H(q_2) - \int\int q_1(x)q_2(y) \left\{ -\frac{1}{2} \ln\ (2\pi) - \frac{(x-\tilde{\mu})^2}{2y} + (\widetilde{\alpha}-\frac{1}{2})\ \ln\ \tilde{\beta} \right. \\ & \left. - \ln\ \Gamma(\widetilde{\alpha}-\frac{1}{2}) - (\tilde{\alpha}+1)\ \ln\ y - \frac{\tilde{\beta}}{y} \right\} dxdy. \end{aligned} \tag{A1}$$

Since, $H(q_1)$ and $H(q_2)$ are Shannon entropy, we have

$$\begin{aligned} -H(q_1) =& \int q_1(x) \ln\ \left(\frac{1}{\sqrt{2\pi\tilde{v}}} \exp\{-\frac{(x-\tilde{\mu})^2}{2\tilde{v}}\} \right) dx \\ =& -\frac{1}{2} \ln\ (2\pi\tilde{v}) - \frac{1}{2}, \end{aligned} \tag{A2}$$

and

$$\begin{aligned} -H(q_2) =& \int q_2(y) \ln\ \left(\frac{\widetilde{\beta}^{\widetilde{\alpha}-\frac{1}{2}}}{\Gamma(\widetilde{\alpha}-\frac{1}{2})} y^{-(\widetilde{\alpha}+\frac{1}{2})} \exp\{-\frac{\widetilde{\beta}}{y}\} \right) dy \\ =& \widetilde{\alpha} - \frac{1}{2} + \ln\ (\widetilde{\beta}\Gamma(\widetilde{\alpha}-\frac{1}{2})) - (\widetilde{\alpha}+\frac{1}{2})\psi_0(\widetilde{\alpha}-\frac{1}{2}). \end{aligned} \tag{A3}$$

Also, we need to know $- < \ln\ p(x,y) >_{q_1q_2}$ as well

$$\begin{aligned} - < \ln\ p(x,y) >_{q_1q_2} =& \int\int q_1(x)q_2(y) \left(\frac{1}{2} \ln\ (2\pi) + \frac{(x-\tilde{\mu})^2}{2y} - (\widetilde{\alpha}-\frac{1}{2})\ \ln\ \tilde{\beta} \right. \\ & \left. + \ln\ \Gamma(\widetilde{\alpha}-\frac{1}{2}) + (\widetilde{\alpha}+1)\ \ln\ y + \frac{\tilde{\beta}}{y} \right) dxdy \\ =& \frac{1}{2} \ln\ (2\pi) - (\widetilde{\alpha}-\frac{1}{2})\ \ln\ \tilde{\beta} + \ln\ \Gamma(\tilde{\alpha}-\frac{1}{2}) \\ & + \int\int q_1(x)q_2(y) \left(\frac{(x-\tilde{\mu})^2}{2y} + (\tilde{\alpha}+1)\ \ln\ y + \frac{\tilde{\beta}}{y} \right) dxdy \\ =& \frac{1}{2} \ln\ (2\pi) - (\widetilde{\alpha}-\frac{1}{2})\ \ln\ \tilde{\beta} + \ln\ \Gamma(\tilde{\alpha}-\frac{1}{2}) \\ & + (\widetilde{\alpha}+1)(\ln\ \widetilde{\beta} - \psi_0(\widetilde{\alpha}-\frac{1}{2})) + \frac{(2\widetilde{\alpha}-1)(\widetilde{v}+2\widetilde{\beta})}{4\widetilde{\beta}} \end{aligned} \tag{A4}$$

Thus, the desire function $KL(\tilde{\boldsymbol{\theta}})$ is

$$KL(\tilde{\boldsymbol{\theta}}) = 2 \ln\ \Gamma(\tilde{\alpha}-\frac{1}{2}) - (2\widetilde{\alpha}+\frac{3}{2})\psi_0(\widetilde{\alpha}-\frac{1}{2}) + \frac{(2\widetilde{\alpha}-1)(\widetilde{v}+2\widetilde{\beta})}{4\widetilde{\beta}} + \widetilde{\alpha} + \frac{5}{2} \ln \widetilde{\beta} - \frac{1}{2} \ln\ \tilde{\boldsymbol{v}} - 1, \tag{A5}$$

where, $\psi_0(\cdot)$ is the polygamma function of order 0, or called digamma function. The gradient expression respect to $\tilde{\boldsymbol{\theta}} = (\widetilde{\alpha}, \widetilde{\beta}, \widetilde{\mu}, \widetilde{v})$ is

$$\nabla KL(\tilde{\boldsymbol{\theta}}) = \left(\frac{\widetilde{v}}{2\widetilde{\beta}} - (2\widetilde{\alpha} + \frac{3}{2})\psi_1(\widetilde{\alpha} - \frac{1}{2}) + 2, \quad -(2\widetilde{\alpha} - 1)\frac{\widetilde{v}}{4\widetilde{\beta}^2} + \frac{5}{2\widetilde{\beta}}, \quad 0, \quad \frac{2\widetilde{\alpha} - 1}{4\widetilde{\beta}} - \frac{1}{2\widetilde{v}} \right), \tag{A6}$$

where, $\psi_1(\cdot)$ is the polygamma function of order 1. We substitute (A6) in the gradient-based and natural gradient formulas, so their algorithms are as follows, respectively

Algorithm 2:

$$\begin{aligned}
\widetilde{\alpha}^{(k+1)} &= \widetilde{\alpha}^{(k)} + \gamma \frac{\partial KL}{\partial \widetilde{\alpha}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{\alpha}^{(k)} - \gamma \left(\frac{\widetilde{v}^{(k)}}{2\widetilde{\beta}^{(k)}} - (2\widetilde{\alpha}^{(k)} + \frac{3}{2})\psi_1(\widetilde{\alpha}^{(k)} - \frac{1}{2}) + 2 \right), \\
\widetilde{\beta}^{(k+1)} &= \widetilde{\beta}^{(k)} - \gamma \frac{\partial KL}{\partial \widetilde{\beta}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{\beta}^{(k)} - \gamma \left(-(2\widetilde{\alpha}^{(k)} - 1)\frac{\widetilde{v}^{(k)}}{4[\widetilde{\beta}^{(k)}]^2} + \frac{5}{2\widetilde{\beta}^{(k)}} \right), \\
\widetilde{\mu}^{(k+1)} &= \widetilde{\mu}^{(k)}, \\
\widetilde{v}^{(k+1)} &= \widetilde{v}^{(k)} - \gamma \frac{\partial KL}{\partial \widetilde{v}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{v}^{(k+1)} - \gamma \left(\frac{2\widetilde{\alpha}^{(k)} - 1}{4\widetilde{\beta}^{(k)}} - \frac{1}{2\widetilde{v}^{(k)}} \right).
\end{aligned}$$

Algorithm 3:

$$\begin{aligned}
\widetilde{\alpha}^{(k+1)} &= \widetilde{\alpha}^{(k)} - \frac{1}{||\Delta KL||} \frac{\partial KL}{\partial \widetilde{\alpha}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{\alpha}^{(k)} - \frac{1}{||\Delta KL||} \left(\frac{\widetilde{v}^{(k)}}{2\widetilde{\beta}^{(k)}} - (2\widetilde{\alpha}^{(k)} + \frac{3}{2})\psi_1(\widetilde{\alpha}^{(k)} - \frac{1}{2}) + 2 \right), \\
\widetilde{\beta}^{(k+1)} &= \widetilde{\beta}^{(k)} - \frac{1}{||\Delta KL||} \frac{\partial KL}{\partial \widetilde{\beta}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{\beta}^{(k)} - \frac{1}{||\Delta KL||} \left(-(2\widetilde{\alpha}^{(k)} - 1)\frac{\widetilde{v}^{(k)}}{4[\widetilde{\beta}^{(k)}]^2} + \frac{5}{2\widetilde{\beta}^{(k)}} \right), \\
\widetilde{\mu}^{(k+1)} &= \widetilde{\mu}^{(k)}, \\
\widetilde{v}^{(k+1)} &= \widetilde{v}^{(k)} - \frac{1}{||\Delta KL||} \frac{\partial KL}{\partial \widetilde{v}}(\widetilde{\alpha}^{(k)}, \widetilde{\beta}^{(k)}, \widetilde{v}^{(k)}) \\
&= \widetilde{v}^{(k)} - \frac{1}{||\Delta KL||} \left(\frac{2\widetilde{\alpha}^{(k)} - 1}{4\widetilde{\beta}^{(k)}} - \frac{1}{2\widetilde{v}^{(k)}} \right).
\end{aligned}$$

Appendix B Conjugate Posterior of Normal Inverse Wishart of $(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}})$

The Normal Inverse Wishart distribution function $\mathcal{NIW}(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}} | \widetilde{\boldsymbol{\mu}}_0, \widetilde{\kappa}, \tilde{\boldsymbol{\Psi}}, \widetilde{v})$ is

$$p(\widetilde{\boldsymbol{\mu}}, \widetilde{\boldsymbol{\Sigma}} | \widetilde{\boldsymbol{\mu}}_0, \widetilde{\kappa}, \tilde{\boldsymbol{\Psi}}, \widetilde{v}) = \mathcal{N}\left(\widetilde{\boldsymbol{\mu}} | \widetilde{\boldsymbol{\mu}}_0, \frac{1}{\widetilde{\kappa}} \widetilde{\boldsymbol{\Sigma}} \right) \mathcal{IW}(\widetilde{\boldsymbol{\Sigma}} | \tilde{\boldsymbol{\Psi}}, \widetilde{v}). \tag{A7}$$

doi:10.20944/preprints202208.0234.v1

The explicit form of joint distribution $p(\boldsymbol{x},\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})$ is

$$\begin{aligned}p(\boldsymbol{x},\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}) =&p(\boldsymbol{x}|\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})p(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}}|\widetilde{\boldsymbol{\mu}}_0,\tilde{\kappa},\tilde{\boldsymbol{\Psi}},\tilde{\nu}) \\ =&\frac{1}{(2\pi)^{\frac{np}{2}}|\widetilde{\boldsymbol{\Sigma}}|^{\frac{n}{2}}}\exp\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\widetilde{\boldsymbol{\mu}})\} \\ &\frac{\sqrt{\tilde{\kappa}}}{(2\pi)^{\frac{p}{2}}|\widetilde{\boldsymbol{\Sigma}}|^{\frac{1}{2}}}\exp\{-\frac{\tilde{\kappa}}{2}\,(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)\}\frac{|\tilde{\boldsymbol{\Psi}}|^{\frac{\tilde{\nu}}{2}}}{2^{\frac{\tilde{\nu}p}{2}}\Gamma_p(\frac{\tilde{\nu}}{2})}|\widetilde{\boldsymbol{\Sigma}}|^{-\frac{\tilde{\nu}+p+1}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Psi}}\widetilde{\boldsymbol{\Sigma}}^{-1}\right]\}.\end{aligned} \tag{A8}$$

We rewrite the exponential expression of multivariate Normal Distribution as follows

$$\sum_{i=1}^{n}(\boldsymbol{x}_i-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{x}_i-\widetilde{\boldsymbol{\mu}})=n\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)^T\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)+\sum_{i=1}^{n}\left(\boldsymbol{x}_i-\overline{\boldsymbol{x}}\right)^T\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\boldsymbol{x}_i-\overline{\boldsymbol{x}}\right). \tag{A9}$$

By substituting (A9) into (A8), we get

$$\begin{aligned}p(\boldsymbol{x},\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})\propto&|\widetilde{\boldsymbol{\Sigma}}|^{-\frac{\tilde{\nu}+p+n+2}{2}}\exp\left\{-\frac{n}{2}\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)^T\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)-\frac{\tilde{\kappa}}{2}\,(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)\right. \\ &\left.-\frac{1}{2}\mathrm{Tr}\left[\left(\tilde{\boldsymbol{\Psi}}+\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}\right]\right\}.\end{aligned} \tag{A10}$$

We have to make integration over $\boldsymbol{x}$ and $\widetilde{\boldsymbol{\Sigma}}$ to get the margin of $\widetilde{\boldsymbol{\mu}}$

$$\begin{aligned}<\ln p(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})>_{q_1q_3}\propto&-\frac{n}{2}<\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)^T\widetilde{\boldsymbol{\Sigma}}^{-1}\left(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}}\right)>_{q_1q_3}-\frac{\tilde{\kappa}}{2}<(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)>_{q_1q_3} \\ \propto&-\frac{\tilde{\kappa}+n}{2}(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n})^T\tilde{\boldsymbol{\Lambda}}^{-1}(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n}),\end{aligned} \tag{A11}$$

where, $<\widetilde{\boldsymbol{\Sigma}}^{-1}>_{q_3}=\tilde{\boldsymbol{\Lambda}}^{-1}$.Thus, $\widetilde{\boldsymbol{\mu}}$ has a Normal distribution $\mathcal{N}\left(\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n},\frac{1}{\tilde{\kappa}+n}\tilde{\boldsymbol{\Lambda}}\right)$. Also, we have the same for the margin of $\widetilde{\boldsymbol{\Sigma}}$

$$\begin{aligned}<\ln p(\widetilde{\boldsymbol{\mu}},\widetilde{\boldsymbol{\Sigma}})>_{q_1q_2}\propto&-\frac{\tilde{\nu}+p+n+2}{2}\ln|\widetilde{\boldsymbol{\Sigma}}|-\frac{n}{2}<(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\overline{\boldsymbol{x}}-\widetilde{\boldsymbol{\mu}})>_{q_1q_2} \\ &-\frac{\tilde{\kappa}}{2}<(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)>_{q_1q_2}-\frac{1}{2}\mathrm{Tr}\left[\left(\tilde{\boldsymbol{\Psi}}+\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}\right] \\ \propto&-\frac{\tilde{\nu}+p+n+2}{2}\ln|\widetilde{\boldsymbol{\Sigma}}|-\frac{\tilde{\kappa}+n}{2}<(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n})^T\widetilde{\boldsymbol{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n})>_{q_1q_2} \\ &-\frac{1}{2}\mathrm{Tr}\left[\left(\tilde{\boldsymbol{\Psi}}+\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}\right] \\ \propto&-\frac{\tilde{\nu}+p+n+2}{2}\ln|\widetilde{\boldsymbol{\Sigma}}|-\frac{1}{2}\mathrm{Tr}\left[\left((\tilde{\kappa}+n)<(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n})(\widetilde{\boldsymbol{\mu}}-\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\boldsymbol{x}}}{\tilde{\kappa}+n})^T>_{q_1q_2}\right.\right. \\ &\left.\left.+\tilde{\boldsymbol{\Psi}}+\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}\right] \\ \propto&-\frac{\tilde{\nu}+p+n+2}{2}\ln|\widetilde{\boldsymbol{\Sigma}}|-\frac{1}{2}\mathrm{Tr}\left[\left(\tilde{\boldsymbol{\Lambda}}+\tilde{\boldsymbol{\Psi}}+\sum_{i=1}^{n}(\boldsymbol{x}_i-\overline{\boldsymbol{x}})(\boldsymbol{x}_i-\overline{\boldsymbol{x}})^T\right)\widetilde{\boldsymbol{\Sigma}}^{-1}\right]\end{aligned} \tag{A12}$$

So, $\widetilde{\mathbf{\Sigma}}$ has the structure of an inverse Wishart distribution with below parameters as following

$$\mathcal{IW}\left(\tilde{\mathbf{\Lambda}}+\tilde{\mathbf{\Psi}}+\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T,\tilde{\nu}+n\right). \tag{A13}$$

It is clear that $\mathbf{x}$ has again a multivariate Normal distribution whose mean and variance-covariance matrix are in the following expression

$$\mathcal{MN}\left(\frac{\tilde{\kappa}\widetilde{\boldsymbol{\mu}}_0+n\overline{\mathbf{x}}}{\tilde{\kappa}+n},\frac{1}{\tilde{\nu}+n-p-1}\left[\tilde{\mathbf{\Lambda}}+\tilde{\mathbf{\Psi}}+\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T\right]\right). \tag{A14}$$

In the next step, we want to make an iterative algorithm of the above mentation marginal distributions. We need some pre-guesses for mean $\widetilde{\boldsymbol{\mu}}_0^{(1)}=\boldsymbol{\mu}_0$, variance $\tilde{\mathbf{\Lambda}}^{(1)}=\mathbf{\Lambda}_0$, and precision $\tilde{\kappa}^{(1)}=\kappa_0$ of $\widetilde{\boldsymbol{\mu}}$ and also for scale matrix $\tilde{\mathbf{\Psi}}^{(1)}=\Psi_0$ and degree of freedom $\tilde{\nu}^{(1)}=\nu_0+n$ of $\widetilde{\mathbf{\Sigma}}$. So, we can consider the alternative algorithm for Normal Inverse Wishart distribution as follows **Algorithm 1:**

$$\begin{aligned}
\tilde{\kappa}^{(k+1)} &= \tilde{\kappa}^{(k)},\\
\widetilde{\boldsymbol{\mu}}_0^{(k+1)} &= \frac{\tilde{\kappa}^{(k)}\widetilde{\boldsymbol{\mu}}_0^{(k)}+n\overline{\mathbf{x}}}{\tilde{\kappa}^{(k)}+n},\\
\tilde{\mathbf{\Lambda}}^{(k+1)} &= \frac{1}{\tilde{\kappa}^{(k)}+n}\tilde{\mathbf{\Lambda}}^{(k)},\\
\tilde{\mathbf{\Psi}}^{(k+1)} &= [\tilde{\mathbf{\Lambda}}^{(k)}]^{-1}+\tilde{\mathbf{\Psi}}^{(k)}+\sum_{i=1}^{n}(\mathbf{x}_i-\overline{\mathbf{x}})(\mathbf{x}_i-\overline{\mathbf{x}})^T,\\
\tilde{\nu}^{(k+1)} &= \tilde{\nu}^{(k)}.
\end{aligned}$$

Appendix C Kullback-Leibler Divergence of Normal Inverse Wishart

First of all, we define $\widetilde{\boldsymbol{\theta}}=(\widetilde{\boldsymbol{\mu}}_0,\tilde{\kappa},\tilde{\mathbf{\Lambda}},\tilde{\mathbf{\Psi}},\tilde{\nu})$, and the corresponding Kullback-Leibler function

$$\begin{aligned}
KL(\tilde{\boldsymbol{\theta}}) &= \int\int q_1(\widetilde{\boldsymbol{\mu}})q_2(\widetilde{\mathbf{\Sigma}})\ln\left(\frac{q_1(\widetilde{\boldsymbol{\mu}})q_2(\widetilde{\mathbf{\Sigma}})}{p(\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{\Sigma}})}\right)d\widetilde{\boldsymbol{\mu}}d\widetilde{\mathbf{\Sigma}} \qquad (A15)\\
&= \int\int q_1(\widetilde{\boldsymbol{\mu}})q_2(\widetilde{\mathbf{\Sigma}})\ln\left(\frac{q_2(\widetilde{\mathbf{\Sigma}})}{p(\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{\Sigma}})}\right)d\widetilde{\boldsymbol{\mu}}d\widetilde{\mathbf{\Sigma}}-H(q_1)\\
&= \int q_2(\widetilde{\mathbf{\Sigma}})\left(\ln q_2(\widetilde{\mathbf{\Sigma}})-\int q_1(\widetilde{\boldsymbol{\mu}})\ln p(\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{\Sigma}})d\widetilde{\boldsymbol{\mu}}\right)d\widetilde{\mathbf{\Sigma}}-H(q_1).
\end{aligned}$$

We need to compute the internal integral first

$$\begin{aligned}
\int q_1(\widetilde{\boldsymbol{\mu}})\ln p(\widetilde{\boldsymbol{\mu}},\widetilde{\mathbf{\Sigma}})d\widetilde{\boldsymbol{\mu}} &= \int q_1(\widetilde{\boldsymbol{\mu}})\ln\left\{\frac{\sqrt{\tilde{\kappa}}|\tilde{\mathbf{\Psi}}|^{\frac{\tilde{\nu}}{2}}}{(2\pi)^{\frac{p}{2}}2^{\frac{\tilde{\nu}p}{2}}\Gamma_p(\frac{\tilde{\nu}}{2})}\right. \qquad (A16)\\
&\qquad \left.|\widetilde{\mathbf{\Sigma}}|^{-\frac{\tilde{\nu}+p+2}{2}}\exp\{-\frac{\tilde{\kappa}}{2}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\mathbf{\Sigma}}^{-1}(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)-\frac{1}{2}\mathrm{Tr}\left[\tilde{\mathbf{\Psi}}\widetilde{\mathbf{\Sigma}}^{-1}\right]\}\right\}d\widetilde{\boldsymbol{\mu}}\\
&= \ln\left\{\frac{\sqrt{\tilde{\kappa}}|\tilde{\mathbf{\Psi}}|^{\frac{\tilde{\nu}}{2}}}{(2\pi)^{\frac{p}{2}}2^{\frac{\tilde{\nu}p}{2}}\Gamma_p(\frac{\tilde{\nu}}{2})}|\widetilde{\mathbf{\Sigma}}|^{-\frac{\tilde{\nu}+p+2}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}\left[\tilde{\mathbf{\Psi}}\widetilde{\mathbf{\Sigma}}^{-1}\right]\}\right\}\\
&\qquad -\frac{\tilde{\kappa}}{2}\int q_1(\widetilde{\boldsymbol{\mu}})\mathrm{Tr}\left[(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)(\widetilde{\boldsymbol{\mu}}-\widetilde{\boldsymbol{\mu}}_0)^T\widetilde{\mathbf{\Sigma}}^{-1}\right]d\widetilde{\boldsymbol{\mu}}\\
&= \ln\left\{\frac{\sqrt{\tilde{\kappa}}|\tilde{\mathbf{\Psi}}|^{\frac{\tilde{\nu}}{2}}}{(2\pi)^{\frac{p}{2}}2^{\frac{\tilde{\nu}p}{2}}\Gamma_p(\frac{\tilde{\nu}}{2})}|\widetilde{\mathbf{\Sigma}}|^{-\frac{\tilde{\nu}+p+2}{2}}\exp\{-\frac{1}{2}\mathrm{Tr}\left[(\tilde{\mathbf{\Lambda}}+\tilde{\mathbf{\Psi}})\widetilde{\mathbf{\Sigma}}^{-1}\right]\}\right\}.
\end{aligned}$$

By substituting (A16) in side of (A15), we get

$$KL(\tilde{\boldsymbol{\theta}}) = \int q_2(\widetilde{\boldsymbol{\Sigma}}) \left(\ln q_2(\widetilde{\boldsymbol{\Sigma}}) - \ln \left\{ \frac{\sqrt{\tilde{\kappa}} |\tilde{\boldsymbol{\Psi}}|^{\frac{\tilde{\nu}}{2}}}{(2\pi)^{\frac{p}{2}} 2^{\frac{\tilde{\nu}p}{2}} \Gamma_p(\frac{\tilde{\nu}}{2})} |\widetilde{\boldsymbol{\Sigma}}|^{-\frac{\tilde{\nu}+p+2}{2}} \exp\{-\frac{1}{2} \mathrm{Tr} \left[(\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}) \widetilde{\boldsymbol{\Sigma}}^{-1} \right] \} \right\} \right) d\widetilde{\boldsymbol{\Sigma}} - H(q_1). \tag{A17}$$

Since $q_2(\widetilde{\boldsymbol{\Sigma}}) = \mathcal{IW}(\tilde{\boldsymbol{\Psi}}, \tilde{\nu})$, we can rewrite $KL(\tilde{\boldsymbol{\theta}})$ as following

$$KL(\tilde{\boldsymbol{\theta}}) = -\ln \left\{ \frac{\sqrt{\tilde{\kappa}} |\tilde{\boldsymbol{\Psi}}|^{\frac{\tilde{\nu}}{2}} \Gamma_p(\frac{\tilde{\nu}+1}{2})}{\pi^{\frac{p}{2}} \Gamma_p(\frac{\tilde{\nu}}{2}) |\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}|^{\frac{\tilde{\nu}+1}{2}}} \right\} + \int q_2(\widetilde{\boldsymbol{\Sigma}}) \ln \left(\frac{\mathcal{IW}(\tilde{\boldsymbol{\Psi}}, \tilde{\nu})}{\mathcal{IW}(\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}, \tilde{\nu} + 1)} \right) d\widetilde{\boldsymbol{\Sigma}} - H(q_1). \tag{A18}$$

The second term is again a Kullback-Leibler function respect to two Inverse Wishart distributions with different parameters calculated in [13]

$$\begin{aligned} \int q_2(\widetilde{\boldsymbol{\Sigma}}) \ln \left(\frac{\mathcal{IW}(\tilde{\boldsymbol{\Psi}}, \tilde{\nu})}{\mathcal{IW}(\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}, \tilde{\nu} + 1)} \right) d\widetilde{\boldsymbol{\Sigma}} = \ln \left(\frac{\Gamma_p(\frac{\tilde{\nu}+1}{2})}{\Gamma_p(\frac{\tilde{\nu}}{2})} \right) + \frac{\tilde{\nu}}{2} \mathrm{Tr} \left[\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I} \right] - \frac{p\tilde{\nu}}{2} \\ - \frac{\tilde{\nu}}{2} \ln |\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I}| - \frac{1}{2} \sum_{i=1}^{p} \psi_0(\frac{\tilde{\nu} - p + i}{2}). \end{aligned} \tag{A19}$$

The last term of (A15) is the Shannon entropy of $\widetilde{\boldsymbol{\mu}}$

$$\begin{aligned} -H(q_1) &= \int q_1(\widetilde{\boldsymbol{\mu}}) \ln q_1(\widetilde{\boldsymbol{\mu}}) d\widetilde{\boldsymbol{\mu}} \\ &= \int q_1(\widetilde{\boldsymbol{\mu}}) \ln \left(\frac{\tilde{\kappa}^{\frac{p}{2}}}{(2\pi)^{\frac{p}{2}} |\tilde{\boldsymbol{\Lambda}}|^{\frac{1}{2}}} \exp\{-\frac{\tilde{\kappa}}{2} (\widetilde{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}_0)^T \tilde{\boldsymbol{\Lambda}}^{-1} (\widetilde{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}_0)\} \right) d\widetilde{\boldsymbol{\mu}} \\ &= \frac{p}{2} \ln(\tilde{\kappa}) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\tilde{\boldsymbol{\Lambda}}| - \frac{\tilde{\kappa}}{2} \int q_1(\widetilde{\boldsymbol{\mu}}) (\widetilde{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}_0)^T \tilde{\boldsymbol{\Lambda}}^{-1} (\widetilde{\boldsymbol{\mu}} - \widetilde{\boldsymbol{\mu}}_0) d\widetilde{\boldsymbol{\mu}} \\ &= \frac{p}{2} \ln(\tilde{\kappa}) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\tilde{\boldsymbol{\Lambda}}| - \frac{1}{2}. \end{aligned} \tag{A20}$$

The final expression of $KL(\tilde{\boldsymbol{\theta}})$ is equivalent to the following after substitution of (A19) and (A20) into (A18)

$$\begin{aligned} KL(\tilde{\boldsymbol{\theta}}) = &- \frac{1}{2} \ln \tilde{\kappa} - \frac{\tilde{\nu}}{2} \ln |\tilde{\boldsymbol{\Psi}}| + \frac{p}{2} \ln \pi - \ln \left(\frac{\Gamma_p(\frac{\tilde{\nu}+1}{2})}{\Gamma_p(\frac{\tilde{\nu}}{2})} \right) + \frac{\tilde{\nu} + 1}{2} \ln |\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}| \\ &+ \ln \left(\frac{\Gamma_p(\frac{\tilde{\nu}+1}{2})}{\Gamma_p(\frac{\tilde{\nu}}{2})} \right) + \frac{\tilde{\nu}}{2} \mathrm{Tr} \left[\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I} \right] - \frac{p\tilde{\nu}}{2} - \frac{\tilde{\nu}}{2} \ln |\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I}| - \frac{1}{2} \sum_{i=1}^{p} \psi_0(\frac{\tilde{\nu} - p + i}{2}) \\ &+ \frac{p}{2} \ln(\tilde{\kappa}) - \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln |\tilde{\boldsymbol{\Lambda}}| - \frac{1}{2} \\ \propto &\frac{p-1}{2} \ln \tilde{\kappa} - \frac{\tilde{\nu}}{2} \ln \left(|\tilde{\boldsymbol{\Psi}}| \cdot |\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I}| \right) + \frac{\tilde{\nu} + 1}{2} \ln |\tilde{\boldsymbol{\Lambda}} + \tilde{\boldsymbol{\Psi}}| \\ &+ \frac{\tilde{\nu}}{2} \mathrm{Tr} \left[\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I} \right] - \frac{p\tilde{\nu}}{2} - \frac{1}{2} \sum_{i=1}^{p} \psi_0(\frac{\tilde{\nu} - p + i}{2}) - \frac{1}{2} \ln |\tilde{\boldsymbol{\Lambda}}| \\ \propto &\frac{p-1}{2} \ln \tilde{\kappa} - \frac{p\tilde{\nu}}{2} + \frac{1}{2} \ln |\mathbf{I} + \tilde{\boldsymbol{\Lambda}}^{-1} \tilde{\boldsymbol{\Psi}}| + \frac{\tilde{\nu}}{2} \mathrm{Tr} \left[\tilde{\boldsymbol{\Psi}}^{-1} \tilde{\boldsymbol{\Lambda}} + \mathbf{I} \right] - \frac{1}{2} \sum_{i=1}^{p} \psi_0(\frac{\tilde{\nu} - p + i}{2}). \end{aligned} \tag{A21}$$

The gradient of $KL(\tilde{\boldsymbol{\theta}})$ respect to $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\mu}}_0, \tilde{\kappa}, \tilde{\boldsymbol{\Lambda}}, \tilde{\boldsymbol{\Psi}}, \tilde{\nu})$ is

$$\begin{aligned}\nabla KL(\tilde{\boldsymbol{\theta}}) = (0,\quad & \frac{p-1}{2\tilde{\kappa}},\quad \frac{1}{2}[\tilde{\boldsymbol{\Lambda}}+\tilde{\boldsymbol{\Psi}}]^{-1} - \frac{1}{2}\tilde{\boldsymbol{\Lambda}}^{-1} + \frac{\tilde{\nu}}{2}\tilde{\boldsymbol{\Psi}}^{-1}, \\ & \frac{1}{2}[\tilde{\boldsymbol{\Lambda}}+\tilde{\boldsymbol{\Psi}}]^{-1} - \frac{\tilde{\nu}}{2}\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}}\tilde{\boldsymbol{\Psi}}^{-1},\quad -\frac{p}{2} + \frac{1}{2}\mathrm{Tr}\left[\tilde{\boldsymbol{\Psi}}^{-1}\tilde{\boldsymbol{\Lambda}} + \mathbf{I}\right] - \frac{1}{4}\sum_{i=1}^{p}\psi_1(\frac{\tilde{\nu}-p+i}{2})\Bigg).\end{aligned} \tag{A22}$$

So, the two second algorithms are

Algorithm 2:

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_0^{(k+1)} &= \widetilde{\boldsymbol{\mu}}_0^{(k)}, \\
\tilde{\kappa}^{(k+1)} &= \tilde{\kappa}^{(k)} - \frac{\gamma}{2}\left(\frac{p-1}{\tilde{\kappa}^{(k)}}\right), \\
\tilde{\boldsymbol{\Lambda}}^{(k+1)} &= \tilde{\boldsymbol{\Lambda}}^{(k)} - \frac{\gamma}{2}\left([\tilde{\boldsymbol{\Lambda}}^{(k)} + \tilde{\boldsymbol{\Psi}}^{(k)}]^{-1} - [\tilde{\boldsymbol{\Lambda}}^{(k)}]^{-1} + \tilde{\nu}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\right), \\
\tilde{\boldsymbol{\Psi}}^{(k+1)} &= \tilde{\boldsymbol{\Psi}}^{(k)} - \frac{\gamma}{2}\left([\tilde{\boldsymbol{\Lambda}}^{(k)} + \tilde{\boldsymbol{\Psi}}^{(k)}]^{-1} - \tilde{\nu}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\tilde{\boldsymbol{\Lambda}}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\right), \\
\tilde{\nu}^{(k+1)} &= \tilde{\nu}^{(k)} - \frac{\gamma}{2}\left(\mathrm{Tr}\left[[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\tilde{\boldsymbol{\Lambda}}^{(k)} + \mathbf{I}\right] - \frac{1}{2}\sum_{i=1}^{p}\psi_1(\frac{\tilde{\nu}^{(k)}-p+i}{2}) - p\right),
\end{aligned}$$

and

Algorithm 3:

$$\begin{aligned}
\widetilde{\boldsymbol{\mu}}_0^{(k+1)} &= \widetilde{\boldsymbol{\mu}}_0^{(k)}, \\
\tilde{\kappa}^{(k+1)} &= \tilde{\kappa}^{(k)} - \frac{1}{2\|\Delta KL\|}\left(\frac{p-1}{2\tilde{\kappa}^{(k)}}\right), \\
\tilde{\boldsymbol{\Lambda}}^{(k+1)} &= \tilde{\boldsymbol{\Lambda}}^{(k)} - \frac{1}{2\|\Delta KL\|}\left([\tilde{\boldsymbol{\Lambda}}^{(k)} + \tilde{\boldsymbol{\Psi}}^{(k)}]^{-1} - [\tilde{\boldsymbol{\Lambda}}^{(k)}]^{-1} + \tilde{\nu}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\right), \\
\tilde{\boldsymbol{\Psi}}^{(k+1)} &= \tilde{\boldsymbol{\Psi}}^{(k)} - \frac{1}{2\|\Delta KL\|}\left([\tilde{\boldsymbol{\Lambda}}^{(k)} + \tilde{\boldsymbol{\Psi}}^{(k)}]^{-1} - \tilde{\nu}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\tilde{\boldsymbol{\Lambda}}^{(k)}[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\right), \\
\tilde{\nu}^{(k+1)} &= \tilde{\nu}^{(k)} - \frac{1}{2\|\Delta KL\|}\left(\mathrm{Tr}\left[[\tilde{\boldsymbol{\Psi}}^{(k)}]^{-1}\tilde{\boldsymbol{\Lambda}}^{(k)} + \mathbf{I}\right] - \frac{1}{2}\sum_{i=1}^{p}\psi_1(\frac{\tilde{\nu}^{(k)}-p+i}{2}) - p\right).
\end{aligned}$$

Appendix D VBA Approximation the Linear Inverse Problem

In this example, the process is pretty the same. First, we have to rewrite $p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}})$ and $\ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}})$

$$\begin{aligned}p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) = & \frac{1}{(2\pi)^{\frac{np}{2}}|\tilde{v}_\epsilon \boldsymbol{I}|^{\frac{n}{2}}}\exp\{-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})^T(\tilde{v}_\epsilon \boldsymbol{I})^{-1}(\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})\} \\ & \frac{1}{(2\pi)^{\frac{p}{2}}|\mathrm{diag}\,[\tilde{\boldsymbol{v}}]\,|^{\frac{1}{2}}}\exp\{-\frac{1}{2}\widetilde{\boldsymbol{f}}^T\mathrm{diag}\,[\tilde{\boldsymbol{v}}]^{-1}\,\widetilde{\boldsymbol{f}}\}\prod_{j=1}^{p}\frac{\widetilde{\beta}_j^{\widetilde{\alpha}_j}}{\Gamma(\widetilde{\alpha}_j)}\tilde{v}_j^{-(\widetilde{\alpha}_j+1)}\exp\{-\frac{\widetilde{\beta}_j}{\tilde{v}_j}\},\end{aligned} \tag{A23}$$

and

$$\begin{aligned}\ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) \propto & -\frac{n}{2}\ln|\tilde{v}_\epsilon \boldsymbol{I}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})^T(\tilde{v}_\epsilon \boldsymbol{I})^{-1}(\boldsymbol{g}_i - \widetilde{\boldsymbol{f}}) - \frac{1}{2}\ln|\mathrm{diag}\,[\tilde{\boldsymbol{v}}]\,| \\ & -\frac{1}{2}\widetilde{\boldsymbol{f}}^T\mathrm{diag}\,[\tilde{\boldsymbol{v}}]^{-1}\,\widetilde{\boldsymbol{f}} + \sum_{j=1}^{p}\widetilde{\alpha}_j\ln\widetilde{\beta}_j - \sum_{j=1}^{p}\ln\Gamma(\widetilde{\alpha}_j) - \sum_{j=1}^{p}(\widetilde{\alpha}_j+1)\ln\tilde{v}_j - \sum_{j=1}^{p}\frac{\widetilde{\beta}_j}{\tilde{v}_j},\end{aligned} \tag{A24}$$

where, $\boldsymbol{I}$ is an identical matrix $p \times p$. According to the above expression, it is easy to find out all margins function starting from $\tilde{\boldsymbol{v}}$

$$\begin{aligned} < \ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) >_{\boldsymbol{g}, \widetilde{\boldsymbol{f}}} &\propto -\frac{1}{2} \ln |\operatorname{diag}[\tilde{\boldsymbol{v}}]| - \frac{1}{2} < \widetilde{\boldsymbol{f}}^{T} \operatorname{diag}[\tilde{\boldsymbol{v}}]^{-1} \widetilde{\boldsymbol{f}} >_{\widetilde{\boldsymbol{f}}} - \sum_{j=1}^{p} (\widetilde{\alpha}_j + 1) \ln \tilde{v}_j - \sum_{j=1}^{p} \frac{\widetilde{\beta}_j}{\tilde{v}_j} \\ &\propto -\frac{1}{2} \sum_{j=1}^{p} \ln \tilde{v}_i - \frac{1}{2} \sum_{j=1}^{p} \frac{< \tilde{f}_k^2 >_{\widetilde{\boldsymbol{f}}}}{\tilde{v}_i} - \sum_{j=1}^{p} (\widetilde{\alpha}_j + 1) \ln \tilde{v}_j - \sum_{j=1}^{p} \frac{\widetilde{\beta}_j}{\tilde{v}_j} \\ &\propto -(\widetilde{\alpha}_k + \frac{3}{2}) \ln \tilde{v}_k - (\frac{\tilde{v}_{\tilde{f}_k} + \widetilde{\mu}^2_{\tilde{f}_k}}{2} + \widetilde{\beta}_k) \frac{1}{\tilde{v}_k}. \end{aligned} \tag{A25}$$

Thus, $\tilde{v}_k \sim \mathcal{IG}(\widetilde{\alpha}_k, \frac{\tilde{v}_k + \widetilde{\mu}^2_{\tilde{f}_k}}{2} + \widetilde{\beta}_k)$ for $k = 1, \cdots, p$. The process for computation of $\boldsymbol{g}$ density function is similar

$$\begin{aligned} < \ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) >_{\widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}} &\propto -\frac{1}{2} \sum_{i=1}^{n} < (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})^T (\tilde{v}_\epsilon \boldsymbol{I})^{-1} (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})_{\widetilde{\boldsymbol{f}}} \\ &= -\frac{1}{2\tilde{v}_\epsilon} \sum_{i=1}^{n} \sum_{j=1}^{p} < (g_{ij} - \tilde{f}_j)^2 >_{\widetilde{\boldsymbol{f}}} \\ &\propto -\frac{1}{2\tilde{v}_\epsilon} \sum_{i=1}^{n} \sum_{j=1}^{p} (g_{ij} - < \tilde{f}_j >_{\widetilde{\boldsymbol{f}}})^2 \\ &= -\frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{g}_i - \widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}})^T (\tilde{v}_\epsilon \boldsymbol{I})^{-1} (\boldsymbol{g}_i - \widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}). \end{aligned} \tag{A26}$$

So, $\boldsymbol{g}$ has again a Normal distribution of $\mathcal{N}(\widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}, \tilde{v}_\epsilon \boldsymbol{I})$. Since, we need the density of $\overline{\boldsymbol{g}}$ in the proceeding, we can specify it now. Therefore, back to properties of Normal distribution, we know that $\overline{\boldsymbol{g}}$ has also a Normal density shown by $\overline{\boldsymbol{g}} \sim \mathcal{N}(\widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}, \frac{1}{n} \tilde{v}_\epsilon \boldsymbol{I})$. Now, we redo all calculation on (A27) for the objective margin of $\widetilde{\boldsymbol{f}}$, separately for each component denoted by $\tilde{f}_k$ for $k = 1, \cdots, p$

$$\begin{aligned} < \ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) >_{\boldsymbol{g}, \tilde{\boldsymbol{v}}} &\propto -\frac{1}{2} \sum_{i=1}^{n} < (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})^T (\tilde{v}_\epsilon \boldsymbol{I})^{-1} (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}}) >_{\boldsymbol{g}} - \frac{1}{2} < \widetilde{\boldsymbol{f}}^{T} \operatorname{diag}[\tilde{\boldsymbol{v}}]^{-1} \widetilde{\boldsymbol{f}} >_{\tilde{\boldsymbol{v}}} \\ &\propto -\frac{1}{2} \left(n < (\overline{\boldsymbol{g}} - \widetilde{\boldsymbol{f}})^T (\tilde{v}_\epsilon \boldsymbol{I})^{-1} (\overline{\boldsymbol{g}} - \widetilde{\boldsymbol{f}}) >_{\boldsymbol{g}} + < \widetilde{\boldsymbol{f}}^{T} \operatorname{diag}[\tilde{\boldsymbol{v}}]^{-1} \widetilde{\boldsymbol{f}} >_{\tilde{\boldsymbol{v}}} \right) \\ &= -\frac{1}{2} \left(\frac{n}{\tilde{v}_\epsilon} \sum_{j=1}^{p} < (\overline{g}_j - \tilde{f}_j)^2 >_{\boldsymbol{g}} + \sum_{j=1}^{p} < \frac{1}{\tilde{v}_j} >_{\tilde{\boldsymbol{v}}} \tilde{f}_j^2 \right) \\ &\propto -\frac{1}{2} \sum_{j=1}^{p} (\frac{n}{\tilde{v}_\epsilon} + < \frac{1}{\tilde{v}_j} >_{\tilde{\boldsymbol{v}}}) \left(\tilde{f}_j - \frac{< \overline{g}_j >_{\boldsymbol{g}}}{1 + \frac{\tilde{v}_\epsilon}{n} < \frac{1}{\tilde{v}_j} >_{\tilde{\boldsymbol{v}}}} \right)^2 \\ &= -\frac{1}{2} \sum_{j=1}^{p} (\frac{n}{\tilde{v}_\epsilon} + \frac{2\widetilde{\alpha}_j}{\tilde{v}_j + \widetilde{\mu}^2_{\tilde{f}_j} + 2\widetilde{\beta}_j}) \left(\tilde{f}_j - \frac{\widetilde{\mu}_{g_i}}{1 + \frac{2\tilde{v}_\epsilon \widetilde{\alpha}_j}{n(\tilde{v}_j + \widetilde{\mu}^2_{\tilde{f}_j} + 2\widetilde{\beta}_j)}} \right)^2. \end{aligned} \tag{A27}$$

Thus, $\tilde{f}_k$ has a Normal distribution with these structures

$$\tilde{f}_k \sim \mathcal{N}\left(\frac{\widetilde{\mu}_{g_k}}{1+\frac{2\tilde{v}_\epsilon\tilde{\alpha}_k}{n(\tilde{v}_k+\widetilde{\mu}^2_{\tilde{f}_k}+2\widetilde{\beta}_k)}}, \frac{\tilde{v}_\epsilon(\tilde{v}_k+\widetilde{\mu}^2_{\tilde{f}_k}+2\widetilde{\beta}_k)}{n(\tilde{v}_k+\widetilde{\mu}^2_{\tilde{f}_k})+2n\widetilde{\beta}_k+2\tilde{v}_\epsilon\tilde{\alpha}_k} \right), \quad k=1,\cdots,p, \tag{A28}$$

so, So, $\widetilde{\boldsymbol{f}}$ is separable due to the diag function. We can summarize all in a multivariate Normal distribution and $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}} = \widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}$

$$\widetilde{\boldsymbol{f}} \sim \mathcal{MN}\left(\frac{\widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}}{1+\frac{2\tilde{v}_\epsilon\tilde{\boldsymbol{\alpha}}}{n(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\tilde{f}}+2\tilde{\boldsymbol{\beta}})}}, \operatorname{diag}\left[\frac{\tilde{v}_\epsilon(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\tilde{f}}+2\tilde{\boldsymbol{\beta}})}{n(\tilde{\boldsymbol{v}}+\widetilde{\boldsymbol{\mu}}^2_{\tilde{f}})+2n\tilde{\boldsymbol{\beta}}+2\tilde{v}_\epsilon\tilde{\boldsymbol{\alpha}}} \right] \right). \tag{A29}$$

where all the mathematical operations on vectors are componentwise, and we get that $\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}} = \widetilde{\boldsymbol{\mu}}_{\widetilde{\boldsymbol{f}}}$. The recursive algorithm is fixed for $\tilde{v}_\epsilon$ and $\tilde{\boldsymbol{\alpha}}$, which can be estimated from the data set. The alternative algorithm is in the following for other parameters

Algorithm 1:

$$\begin{aligned}
\tilde{v}_\epsilon{}^{(k+1)} &= \tilde{v}_\epsilon{}^{(k)}, \\
\tilde{\boldsymbol{\alpha}}^{(k+1)} &= \tilde{\boldsymbol{\alpha}}^{(k)}, \\
\tilde{\boldsymbol{\beta}}^{(k+1)} &= \frac{1}{2}(\tilde{\boldsymbol{v}}^{(k)} + [\widetilde{\boldsymbol{\mu}}^{(k)}_{\widetilde{\boldsymbol{f}}}]^2) + \tilde{\boldsymbol{\beta}}^{(k)}, \\
\widetilde{\boldsymbol{\mu}}^{(k+1)}_{\widetilde{\boldsymbol{f}}} &= \widetilde{\boldsymbol{\mu}}^{(k)}_{\widetilde{\boldsymbol{f}}} \left(1 + \frac{2\tilde{v}_\epsilon{}^{(k)}\tilde{\boldsymbol{\alpha}}^{(k)}}{n(\tilde{\boldsymbol{v}}^{(k)} + [\widetilde{\boldsymbol{\mu}}^{(k)}_{\widetilde{\boldsymbol{f}}}]^2 + 2\tilde{\boldsymbol{\beta}}^{(k)})} \right)^{-1}, \\
\tilde{\boldsymbol{v}}^{(k+1)} &= \operatorname{diag}\left[\frac{\tilde{v}_\epsilon{}^{(k)}(\tilde{\boldsymbol{v}}^{(k)} + [\widetilde{\boldsymbol{\mu}}^{(k)}_{\widetilde{\boldsymbol{f}}}]^2 + 2\tilde{\boldsymbol{\beta}}^{(k)})}{n(\tilde{\boldsymbol{v}}^{(k)} + [\widetilde{\boldsymbol{\mu}}^{(k)}_{\widetilde{\boldsymbol{f}}}]^2) + 2n\tilde{\boldsymbol{\beta}}^{(k)} + 2\tilde{v}_\epsilon{}^{(k)}\tilde{\boldsymbol{\alpha}}^{(k)}} \right].
\end{aligned}$$

Appendix E $KL(\widetilde{\boldsymbol{\theta}})$ the Linear Inverse Problem

We define $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}, \tilde{v}_\epsilon, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$ and its corresponding $KL(\widetilde{\boldsymbol{\theta}})$

$$KL(\widetilde{\boldsymbol{\theta}}) = -H(q_1) - H(q_2) - H(q_3) - < \ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}}) >_{q_1 q_2 q_3}. \tag{A30}$$

We separately decompose each term initializing from $-H(q_1)$

$$\begin{aligned}
-H(q_1) &= \int q_1(\boldsymbol{g}) \ln \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\operatorname{diag}[\tilde{v}_\epsilon \mathcal{I}]|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\boldsymbol{g} - \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}})^T \operatorname{diag}[\tilde{\boldsymbol{v}}]^{-1} (\boldsymbol{g} - \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}})\} \right) d\boldsymbol{g} \\
&= -\frac{p}{2} - \frac{p}{2}\ln(2\pi) - \frac{p}{2}\ln \tilde{v}_\epsilon,
\end{aligned} \tag{A31}$$

and

$$\begin{aligned}
-H(q_2) &= \int q_2(\widetilde{\boldsymbol{f}}) \ln \left(\frac{1}{(2\pi)^{\frac{p}{2}} |\operatorname{diag}[\tilde{\boldsymbol{v}}]|^{\frac{1}{2}}} \exp\{-\frac{1}{2}\widetilde{\boldsymbol{f}}^T \operatorname{diag}[\tilde{\boldsymbol{v}}]^{-1} \widetilde{\boldsymbol{f}}\} \right) d\widetilde{\boldsymbol{f}} \\
&= -\frac{p}{2} - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\operatorname{diag}[\tilde{\boldsymbol{v}}]|,
\end{aligned} \tag{A32}$$

and

$$\begin{aligned} -H(q_3) &= \int q_3(\tilde{v}) \ln \left(\prod_{j=1}^{p} \frac{\widetilde{\beta}_j^{\widetilde{\alpha}_j}}{\Gamma(\widetilde{\alpha}_j)} \tilde{v}_j^{-(\widetilde{\alpha}_j+1)} \exp\{-\frac{\widetilde{\beta}_j}{\tilde{v}_j}\} \right) d\tilde{\boldsymbol{v}} \\ &= -\sum_{j=1}^{p} \widetilde{\alpha}_j - \sum_{j=1}^{p} \ln(\widetilde{\beta}_j \Gamma(\widetilde{\alpha}_j)) + \sum_{j=1}^{p} (\widetilde{\alpha}_j + 1)\psi_0(\widetilde{\alpha}_j), \end{aligned} \tag{A33}$$

and

$$\begin{aligned} -<\ln p(\boldsymbol{g}, \widetilde{\boldsymbol{f}}, \tilde{\boldsymbol{v}})>_{q_1q_2q_3} &\propto \int\int\int q_1(\boldsymbol{g}) q_2(\widetilde{\boldsymbol{f}}) q_3(\tilde{\boldsymbol{v}}) \left(\frac{n}{2} \ln|\tilde{v}_\epsilon \boldsymbol{I}| + \frac{1}{2} \sum_{i=1}^{n} (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}})^T (\tilde{v}_\epsilon \boldsymbol{I})^{-1} (\boldsymbol{g}_i - \widetilde{\boldsymbol{f}}) \right. \\ & + \frac{1}{2} \ln|\mathrm{diag}\,[\tilde{\boldsymbol{v}}]\,| + \frac{1}{2} \widetilde{\boldsymbol{f}}^T \mathrm{diag}\,[\tilde{\boldsymbol{v}}]^{-1} \widetilde{\boldsymbol{f}} - \sum_{j=1}^{p} \widetilde{\alpha}_j \ln \widetilde{\beta}_j \\ & \left. + \sum_{j=1}^{p} \ln\Gamma(\widetilde{\alpha}_j) + \sum_{j=1}^{p} (\widetilde{\alpha}_j + 1) \ln \tilde{v}_j + \sum_{j=1}^{p} \frac{\widetilde{\beta}_j}{\tilde{v}_j} \right) d\boldsymbol{g} d\widetilde{\boldsymbol{f}} d\tilde{\boldsymbol{v}} \\ & \propto \frac{np}{2} \ln(\tilde{v}_\epsilon) + \frac{1}{2\tilde{v}_\epsilon} \sum_{i=1}^{n} \sum_{j=1}^{p} \int\int q_1(\boldsymbol{g}) q_2(\widetilde{\boldsymbol{f}}) (g_{ij}^2 + \tilde{f}_j^2) d\boldsymbol{g} d\widetilde{\boldsymbol{f}} \\ & - \sum_{j=1}^{p} \widetilde{\alpha}_j \ln \widetilde{\beta}_j + \sum_{j=1}^{p} \ln\Gamma(\widetilde{\alpha}_j) + \int q_3(\tilde{\boldsymbol{v}}) \left(\sum_{j=1}^{p} (\widetilde{\alpha}_j + \frac{3}{2}) \ln \tilde{v}_j + \sum_{j=1}^{p} \frac{\widetilde{\beta}_j}{\tilde{v}_j} \right) d\tilde{\boldsymbol{v}} \\ & \propto \frac{np}{2} \ln(\tilde{v}_\epsilon) + \frac{n}{2\tilde{v}_\epsilon} \sum_{j=1}^{p} (2\tilde{v}_j + \widetilde{\mu}_{g_j}^2) - \sum_{j=1}^{p} \widetilde{\alpha}_j \ln \widetilde{\beta}_j + \sum_{j=1}^{p} \ln\Gamma(\widetilde{\alpha}_j) \\ & + \sum_{j=1}^{p} (\widetilde{\alpha}_j + \frac{3}{2})(\ln \widetilde{\beta}_j - \psi_0(\widetilde{\alpha}_j)) + \sum_{j=1}^{p} \widetilde{\alpha}_j. \end{aligned} \tag{A34}$$

Finally, we get

$$\begin{aligned} KL(\widetilde{\boldsymbol{\theta}}) \propto & -\frac{p}{2} \ln \tilde{v}_\epsilon - \frac{1}{2} \ln|\mathrm{diag}\,[\tilde{\boldsymbol{v}}]\,| - \sum_{j=1}^{p} \ln(\widetilde{\beta}_j \Gamma(\widetilde{\alpha}_j)) + \sum_{j=1}^{p} (\widetilde{\alpha}_j + 1)\psi_0(\widetilde{\alpha}_j) + \frac{np}{2} \ln(\tilde{v}_\epsilon) \\ & + \frac{n}{2\tilde{v}_\epsilon} \sum_{j=1}^{p} (2\tilde{v}_j + \widetilde{\mu}_{g_j}^2) - \sum_{j=1}^{p} \widetilde{\alpha}_j \ln \widetilde{\beta}_j + \sum_{j=1}^{p} \ln\Gamma(\widetilde{\alpha}_j) + \sum_{j=1}^{p} (\widetilde{\alpha}_j + \frac{3}{2})(\ln \widetilde{\beta}_j - \psi_0(\widetilde{\alpha}_j)). \\ & \propto \frac{p(n-1)}{2} \ln \tilde{v}_\epsilon + \frac{n}{2\tilde{v}_\epsilon} \sum_{j=1}^{p} (2\tilde{v}_j + \widetilde{\mu}_{g_j}^2) - \frac{1}{2} \sum_{j=1}^{p} \ln \tilde{v}_j - \frac{1}{2} \sum_{j=1}^{p} \psi_0(\widetilde{\alpha}_j) + \frac{1}{2} \sum_{j=1}^{p} \ln \widetilde{\beta}_j. \end{aligned} \tag{A35}$$

To make the last two algorithms, we have to differentiate $KL(\widetilde{\boldsymbol{\theta}})$ respect to $\widetilde{\boldsymbol{\theta}} = (\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}, \tilde{v}_\epsilon, \tilde{\boldsymbol{v}}, \tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\beta}})$

$$\nabla KL(\widetilde{\boldsymbol{\theta}}) = \left(\frac{n}{\tilde{v}_\epsilon} \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}, \quad \frac{p(n-1)}{2\tilde{v}_\epsilon} - \frac{n}{2\tilde{v}_\epsilon{}^2} \sum_{j=1}^{p} (2\tilde{v}_j + \widetilde{\boldsymbol{\mu}}_{g_j}^2), \quad \frac{n}{\tilde{v}_\epsilon} \mathbf{1} - \frac{1}{2} \tilde{\boldsymbol{v}}^{-1}, \quad -\frac{1}{2} \psi_1(\tilde{\boldsymbol{\alpha}}), \quad \frac{1}{2} \tilde{\boldsymbol{\beta}}^{-1} \right), \tag{A36}$$

where **1** is the all-ones vector, and also all operations are componentwise. The gradient and natural gradient algorithms are as follows, respectively

$$\begin{array}{ll}
\textbf{Algorithm2}: & \textbf{Algorithm3}: \\
\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}^{(k+1)} = \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}^{(k)} - \gamma\left(\frac{n}{\tilde{v}_\epsilon}\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}\right), & \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}^{(k+1)} = \widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}^{(k)} - \frac{1}{\|\Delta KL\|}\left(\frac{n}{\tilde{v}_\epsilon}\widetilde{\boldsymbol{\mu}}_{\boldsymbol{g}}\right), \\
\tilde{v_\epsilon}^{(k+1)} = \tilde{v_\epsilon}^{(k)} - \gamma\left(\frac{p(n-1)}{2\tilde{v}_\epsilon} - \frac{n}{2\tilde{v_\epsilon}^2}\sum_{j=1}^{p}(2\tilde{v}_j + \widetilde{\mu}_{g_j}^2)\right), & \tilde{v_\epsilon}^{(k+1)} = \tilde{v_\epsilon}^{(k)} - \frac{1}{\|\Delta KL\|}\left(\frac{p(n-1)}{2\tilde{v}_\epsilon} - \frac{n}{2\tilde{v_\epsilon}^2}\sum_{j=1}^{p}(2\tilde{v}_j + \widetilde{\mu}_{g_j}^2)\right), \\
\tilde{\boldsymbol{v}}^{(k+1)} = \tilde{\boldsymbol{v}}^{(k)} - \gamma\left(\frac{n}{\tilde{v}_\epsilon}\mathbf{1} - \frac{1}{2}\tilde{\boldsymbol{v}}^{-1}\right), & \tilde{\boldsymbol{v}}^{(k+1)} = \tilde{\boldsymbol{v}}^{(k)} - \frac{1}{\|\Delta KL\|}\left(\frac{n}{\tilde{v}_\epsilon}\mathbf{1} - \frac{1}{2}\tilde{\boldsymbol{v}}^{-1}\right), \\
\tilde{\boldsymbol{\alpha}}^{(k+1)} = \tilde{\boldsymbol{\alpha}}^{(k)} - \gamma\left(-\frac{1}{2}\psi_1(\tilde{\boldsymbol{\alpha}})\right), & \tilde{\boldsymbol{\alpha}}^{(k+1)} = \tilde{\boldsymbol{\alpha}}^{(k)} - \frac{1}{\|\Delta KL\|}\left(-\frac{1}{2}\psi_1(\tilde{\boldsymbol{\alpha}})\right), \\
\tilde{\boldsymbol{\beta}}^{(k+1)} = \tilde{\boldsymbol{\beta}}^{(k)} - \gamma\left(\frac{1}{2}\tilde{\boldsymbol{\beta}}^{-1}\right), & \tilde{\boldsymbol{\beta}}^{(k+1)} = \tilde{\boldsymbol{\beta}}^{(k)} - \frac{1}{\|\Delta KL\|}\left(\frac{1}{2}\tilde{\boldsymbol{\beta}}^{-1}\right).
\end{array}$$

The objective of the inverse problem is to approximate the distribution of $\boldsymbol{f}$. We use the distribution of $\boldsymbol{g}$ in fig 3 to show the number of method accuracies. Here are the conjectures of the VBA, gradient-based, and natural gradient algorithms, respectively

$$\begin{cases} \mu_{\widetilde{f}} = \begin{bmatrix} 0.025537 \\ 0.069324 \end{bmatrix}, \\ \widetilde{\Sigma}_{\widetilde{f}} = \begin{bmatrix} 0.165003 & 0 \\ 0 & 0.147284 \end{bmatrix}, \end{cases} \begin{cases} \mu_{\widetilde{f}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ \widetilde{\Sigma}_{\widetilde{f}} = \begin{bmatrix} 0.158464 & 0 \\ 0 & 2.410353 \end{bmatrix}, \end{cases} \begin{cases} \mu_{\widetilde{f}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \\ \widetilde{\Sigma}_{\widetilde{f}} = \begin{bmatrix} 1.559747 & 0 \\ 0 & 1.291366 \end{bmatrix}. \end{cases}$$