A SUBSPACE-BASED VARIATIONAL BAYESIAN METHOD

Yuling ZHENG, Aurélia FRAYSSE, Thomas RODET

L2S, University of Paris-Sud, CNRS, Supelec
3 rue Joliot-Curie, 91192 Gif-sur-Yvette cedex, France
{zheng,fraysse,rodet}@lss.supelec.fr

ABSTRACT

This paper is devoted to an improved variational Bayesian method. Actually, variational Bayesian issue can be seen as a convex functional optimization problem. Our main contribution is the adaptation of subspace optimization methods into the functional space involved in this problem. We highlight the efficiency of our methodology on a linear inverse problem with a sparse prior. Comparisons with classical Bayesian methods through a numerical example show the notable improved computation time.

***Index Terms*—** variational Bayesian, subspace optimization, sparse prior.

1. INTRODUCTION

Bayesian inference is a commonly used methodology for ill-posed inverse problems in signal and image processing. Its objective is to estimate unknown parameters from a posterior distribution which is derived from prior information and the information coming from the data thanks to Bayes' rule. Nevertheless, this posterior distribution depends closely on the partition function which is generally unknown and therefore cannot be explicitly determined. So the main challenge is to retrieve this posterior distribution.

In this context, two main types of methods are employed, stochastic and analytic approximations. The first one is based on Monte Carlo Markov Chains (MCMC) [1], where samples of the desired distribution are simulated. However this method is notoriously numerically expensive. Therefore MacKay in [2], see also [3] for a survey, proposed variational Bayesian approach (VBA) aiming to determine analytic approximations of the posterior law. In this case, the objective is to find a simpler probability density function (pdf) close to the posterior law. This problem is thus formulated as a convex infinite-dimensional optimization problem, whose resolution results in an analytic approximation of the true posterior distribution. However, this approximation has no explicit form and has to be approximated by iterative methods, such as the Gauss Seidel one which is known to be time consuming. The classical Bayesian methodology is thus not efficient, especially when dealing with large dimensional problems. Recently, a different method has been introduced in [4] in order to improve variational Bayesian methodology. The main idea is to transpose a classical optimization method, the gradient descent, into the space of pdf. Mathematical details and convergence proof can be found in [5].

A natural idea to improve the method of [4] is to consider a new descent direction. The first possible choice would be to integrate conjugate gradient methods which converge faster than the gradient descent due to the introduction of memory in the direction. Nevertheless, in our context, the space involved in the optimization problem is no longer a Hilbert space and there is no notion of conjugate directions.

The main contribution of this paper is to adapt the subspace optimization methods [6, 7, 8] to the functional space involved in variational Bayesian approach. The advantage of subspace optimization is its generalized descent directions where Hilbert structure is no longer required. This optimization method is based on descent directions which belong to a subspace of dimension greater than one. This gives more flexibility in the choice of direction and of algorithm step-size, as a result, subspace optimization methods are more efficient than conjugate gradient methods [8].

The rest of this paper is organized as follows: in Section 2, we formulate the optimization problem involved; in Section 3, we present our subspace based variational Bayesian method. Section 4 is devoted to the application to a linear inverse problem whereas in Section 5 simulation results on a small tomographic problem are given together with a numerical comparison with classical methods in terms of reconstruction performances and computation time. Finally Section 6 concludes the paper.

2. STATEMENT OF THE PROBLEM

In the following, $\mathbf{y} \in \mathbb{R}^M$ and $\mathbf{w} \in \mathbb{R}^P$ denote respectively the data vector and the unknown parameter vector to be estimated whereas $p(\mathbf{w})$, $p(\mathbf{w}|\mathbf{y})$ and $q(\mathbf{w})$ represent respectively the prior distribution, the true posterior law and its approximation.

Variational Bayesian approaches assume that $q(\mathbf{w})$ is separable. It could be either an assumption of full separability respectively to all the elements contained in $\mathbf{w}$, or a partial

one, e.g. where only the separability between the unknown variables and the hidden ones is considered. The optimal approximation is obtained by minimizing the Kullback-Leibler ($\mathcal{KL}$) divergence to the true posterior distribution. However, the direct computation of $\mathcal{KL}$ divergence is intractable since it depends on the posterior distribution. But as illustrated in [4, 9], minimizing $\mathcal{KL}$ divergence is equivalent to maximizing the negative free energy $\mathcal{F}(q(\mathbf{w}))$ which depends on the joint distribution $p(\mathbf{y}, \mathbf{w})$. The negative free energy is defined as follows:

$$\mathcal{F}(q(\mathbf{w})) = \int_{\mathbb{R}^N} q(\mathbf{w}) \log \frac{p(\mathbf{y}, \mathbf{w})}{q(\mathbf{w})} d\mathbf{w}. \quad (1)$$

It is therefore calculable and used hence as an alternative to the $\mathcal{KL}$ divergence.

The variational Bayesian problem could be formulated as:

$$q^{opt} = \arg\max_q \mathcal{F}(q(\mathbf{w})), \quad s.t. \ \ q \text{ is a separable p.d.f.} \quad (2)$$

Note that as the space of pdf is not an Hilbert space, classical optimization methods are not directly applicable to this problem. Problem (2) is thus solved in [4] by an approach based on the exponentiated gradient method [10], where at each iteration the approximate distribution q is obtained as a product of the previous iterate and the exponential of the gradient.

To obtain a more efficient method, we have transposed the subspace optimization method into the resolution of our functional optimization problem as in [4].

3. OUR PROPOSED METHOD

Let us firstly describe the subspace optimization algorithms [6, 8]. The iteration count is hereafter denoted by $k \in \mathbb{N}$. Generally speaking, subspace optimization algorithms in a Hilbert space use the following iteration formula:

$$\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k = \mathbf{x}^k + \mathbf{D}^k \mathbf{s}^k, \quad (3)$$

where $\mathbf{d}^k$ is the considered descent direction which is a combination of several directions. Here $\mathbf{D}^k$ gathers a set of I directions spanning the subspace, with I larger than one. Moreover, the vector $\mathbf{s}^k$ denotes the step-size along each direction contained in $\mathbf{D}^k$. In [6] a subspace spanned by the opposite gradient and the previous direction, i.e. $\mathbf{D}^k = [-\mathbf{g}^k, \mathbf{d}^{k-1}]$, has been used. Chouzenoux *et al.* [8] have addressed a discussion about constructions of the subspace as well as the dimension of the subspace through simulation results on several image restoration problems. It has been shown that the subspace constructed by the gradient and the memory to one previous direction gives the best performance in terms of implementation complexity and computation time. Consequently, we have adopted this type of subspace.

As discussed in Section 2, we consider the optimization structure constructed in [4],

$$q^{k+1}(\mathbf{w}) = q^k(\mathbf{w}) h^k(\mathbf{w}), \quad (4)$$

where $h^k \in L^1(q^k)$ is a positive function. The simplest way to obtain a pdf is to impose an exponential structure for h^k, therefore,

$$h^k(\mathbf{w}) = \exp(d^k) = \exp(\mathbf{D}^k \mathbf{s}^k), \quad (5)$$

with

$$\mathbf{D}^k = [\mathrm{d}f(q^k, \mathbf{w}), d^{k-1}], \quad (6)$$

where $\mathrm{d}f(q^k, \mathbf{w})$ is a term obtained from the Gateaux differential of the negative free energy $\mathcal{F}$ at kth estimate q^k, see [5] for details.

The dimension of the subspace decides that the multi-dimensional step is of dimension two: $\mathbf{s} = [s_1, s_2]^T$. We can derive a new estimate depending on the step-size $\mathbf{s}$ by using (5), (6) and the hypothesis of separability for q,

$$\begin{aligned} q^{\mathbf{s}}(\mathbf{w}) &= q^k(\mathbf{w}) \exp(s_1 \mathrm{d}f(q^k, \mathbf{w}) + s_2 d^{k-1}) \\ &= K^k \prod_i q_i^k \left(\frac{\exp\left(\langle \log p(\mathbf{y}, \mathbf{w}) \rangle_{\prod_{j \neq i} q_j^k} \right)}{q_i^k} \right)^{s_1} \left(\frac{q_i^k}{q_i^{k-1}} \right)^{s_2} \end{aligned} \quad (7)$$

where K^k is a normalization constant depending on the step-size $\mathbf{s}$ and $\langle \cdot \rangle_q = \mathbb{E}_q[\cdot]$. We introduce here an auxiliary function as follows:

$$q_i^r = \exp\left(\langle \log p(\mathbf{y}, \mathbf{w}) \rangle_{\prod_{j \neq i} q_j^k} \right). \quad (8)$$

The distribution $q^{\mathbf{s}}(\mathbf{w})$ could therefore be rewritten as

$$q^{\mathbf{s}}(\mathbf{w}) = K^k \prod_i q_i^k \left(\frac{q_i^r}{q_i^k} \right)^{s_1} \left(\frac{q_i^k}{q_i^{k-1}} \right)^{s_2}. \quad (9)$$

The determination of the step-size is a crucial and delicate task. Let us define $f^k(\mathbf{s}) = \mathcal{F}(q^k \exp(\mathbf{D}^k \mathbf{s}))$, the optimal step-size is then defined by

$$(\mathbf{s}^{opt})^k = \arg\max_{\mathbf{s} \in \mathbb{R}^2} f^k(\mathbf{s}). \quad (10)$$

The determination of this optimal step-size is generally expensive. Therefore, in this work, we have defined a suboptimal step-size. Firstly, the second order Taylor expansion of $f^k(\mathbf{s})$ at origin is taken as its local approximation,

$$\tilde{f}^k(\mathbf{s}) = f^k(\mathbf{0}) + \left(\frac{\partial f^k}{\partial \mathbf{s}} \Big|_{\mathbf{s}=\mathbf{0}} \right)^T \mathbf{s} + \frac{1}{2} \mathbf{s}^T \left(\mathbf{H}^k|_{\mathbf{s}=\mathbf{0}} \right) \mathbf{s}, \quad (11)$$

where $\frac{\partial f^k}{\partial \mathbf{s}}$ and $\mathbf{H}^k$ denote respectively the gradient vector and the Hessian matrix with $\mathbf{H}_{ij}^k = \frac{\partial^2 f^k}{\partial s_i \partial s_j}, i, j = 1, 2$.

Then we take the suboptimal step-size that maximizes the quadratic approximation $\tilde{f}^k(\mathbf{s})$, whose maximum is achieved at the point where its derivative vanishes. We have therefore,

$$(\mathbf{s}^{subopt})^k = - \left(\mathbf{H}^k|_{\mathbf{s}=\mathbf{0}} \right)^{-1} \frac{\partial f^k}{\partial \mathbf{s}} \Big|_{\mathbf{s}=\mathbf{0}}. \quad (12)$$

We then obtain the new estimation by substituting the suboptimal step-size into (9).

The first advantage of our proposed method is that updating all the distributions $(q_i)_{i=1,\ldots,P}$ is performed in parallel which leads to a significant acceleration compared to the classical variational Bayesian method. Secondly, it exploits a generalized direction optimization algorithm. At each iteration, an optimization of the direction in a subspace is performed. As a result, our approach is more efficient than the gradient based variational Bayesian approach.

4. APPLICATION TO INVERSE PROBLEM

In this section, we consider the application of the method developed in Section 3 to an ill-posed inverse problem with a sparse prior.

4.1. Posterior distribution

A classical linear forward model is considered:

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\epsilon}, \tag{13}$$

where $\mathbf{x} \in \mathbb{R}^N$ denotes the unknown parameters to be estimated. $\mathbf{A}$ is a known matrix of dimension $M \times N$ and $\boldsymbol{\epsilon}$ is the white Gaussian noise, i.e. $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{\epsilon}}^2\mathbf{I})$.

We take sparse prior into account by considering a separable Student-*t* distribution [11]. Student-*t* distribution depends on a shape parameter and small values of this parameter give heavy-tailed probability density functions. Furthermore, it can be written as a Vector Gaussian Scale Mixture, see [12, 13] which has a Gaussian structure with an inverse variance given by hidden variables $(z_i)_{i=1,\ldots,N}$ of Gamma distribution. Hence, the prior distribution is written as

$$\begin{aligned} p(x_i) &= \int_{\mathbb{R}} \mathcal{N}(x_i|0, \sigma_p^2/z_i)\mathcal{G}(z_i|\nu/2, \nu/2)\mathrm{d}z_i \\ &\propto \int_{\mathbb{R}} \frac{\sqrt{z_i}}{\sqrt{2\pi}\sigma_p} e^{\frac{-z_i x_i^2}{2\sigma_p^2}} z_i^{\nu/2-1} e^{-z_i\nu/2} dz_i, \end{aligned} \tag{14}$$

where $\mathcal{N}$ and $\mathcal{G}$ respectively denote the Normal and Gamma distribution.

According to the considered model, the posterior distribution is thus given by:

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\mathbf{y}) &\propto \sigma_{\boldsymbol{\epsilon}}^{-M} \exp\left[-\frac{\|\mathbf{y} - \mathbf{A}\mathbf{x}\|^2}{2\sigma_{\boldsymbol{\epsilon}}^2}\right] \\ &\times \prod_{i=1}^{N} \frac{\sqrt{z_i}}{\sigma_p} \exp\left[-\frac{z_i x_i^2}{2\sigma_p^2}\right] z_i^{\nu/2-1} e^{-z_i\nu/2}. \end{aligned} \tag{15}$$

We can see from (15) that the advantage of Student-*t* prior is to give conjugate laws, e.g. $p(\mathbf{x}|\mathbf{z})$ is Gaussian which is conjugate for the Gaussian likelihood $p(\mathbf{y}|\mathbf{x}, \mathbf{z})$. Consequently, the posterior law of $\mathbf{x}$ is still Gaussian. This conjugacy is needed in the development of efficient VBA.

4.2. Variational Bayesian algorithm

In this section, we apply the method proposed in Section 3 to the model given in Section 4.1. Since the major objective of this application is to illustrate the efficiency advantage of our method, we only derive the supervised algorithm, but the extension to unsupervised one is possible and will be done in a future work.

In this case, the unknowns are $\mathbf{w} = (\mathbf{x}, \mathbf{z})$. Let us assume that

$$q(\mathbf{x}, \mathbf{z}) = \prod_i q_i(x_i) \prod_j \tilde{q}_j(z_j). \tag{16}$$

Due to the use of the conjugate priors, variational Bayesian algorithms yield a Gaussian distribution of mean $\mathbf{m}_k(i)$ and variance $\boldsymbol{\sigma}_k^2(i)$ for $q_i^k(x_i)_{i=1,\ldots,N}$, and a Gamma distribution with shape and rate parameters denoted by $\mathbf{a}_k(j)$ and $\mathbf{b}_k(j)$ for $\tilde{q}_j(z_j)_{j=1,\ldots,N}$. As a result, updating the distributions $(q_i^k)_{i=1,\ldots,N}$ and $(\tilde{q}_j^k)_{j=1,\ldots,N}$ is performed by updating their parameters.

One can see from (15) that the conditional posterior $p(\mathbf{z}|\mathbf{x}, \mathbf{y})$ is separable so that the classical variational Bayesian approach [3] leads to an explicit solution for $(\tilde{q}_j^{k+1})_{j=1,\ldots,N}$, as shown in [4, 11]. Concerning $\mathbf{x}$, the posterior distribution is more intricate. In this case, we prefer adopting our proposed subspace-based method rather than using the classical variational Bayesian approach. The updating equation defined by (9) is then used. Altogether, our problem could be solved by using the following algorithm.

Algorithm 1 Proposed algorithm

1. Initialize $(q_i^0)_{i=1,\ldots,N}$ and $(\tilde{q}_j^0)_{j=1,\ldots,N}$
2. Update the parameters of Gamma distributions $(\tilde{q}_j^{k+1})_{j=1,\ldots,N}$ with classical VBA
3. Compute auxiliary functions $(q_i^r)_{i=1,\ldots,N}$ using Eq. (8)
4. Determine the *subspace*: $\left[\frac{q^r}{q^k}, \frac{q^k}{q^{k-1}}\right]$
5. Compute the step-size using Eq. (12)
6. Update means and variances of $(q_i^{k+1})_{i=1,\ldots,N}$ using Eq. (9)
7. Go back to 2 until convergence

In Algorithm 1, Step 3 to Step 6 are devoted to update the parameters of the Gaussian distribution.

5. EXPERIMENTAL RESULTS

Our proposed algorithm is evaluated through a comparison with the MCMC approach, classical VBA with hypothesis of full separability (VBFS) as (16), classical VBA with the partial separability only between $\mathbf{x}$ and $\mathbf{z}$ (VBPS) and the gradient-like variational Bayesian approach (Grad) proposed

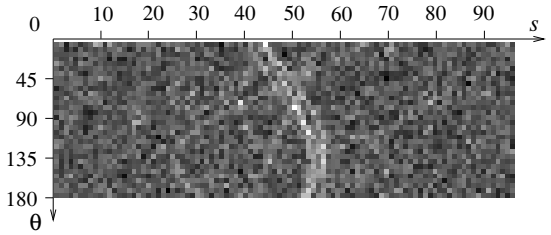

Fig. 1: Data: sinogram composed of 32 projections using 95 detector cells.

in [4]. These five approaches are applied to a tomographic problem with synthetic test data.

The synthetic data (see Fig. 1) is generated from a small test image of dimension 64×64 (see Fig. 2(a)), which is composed of 7 peaks with magnitudes ranging from 0.5 to 1. The data used in the simulation are the projections with a parallel-beam geometry at 32 angles uniformly distributed on $[0, \pi[$. Each projection is collected by 95 detection cells. A Gaussian noise with standard deviation equal to 0.3 is also added. Thus the data has a relatively low signal-to-noise ratio (see Fig. 1). The number of unknowns ($64 \times 64 = 4096$) is much larger than the number of data ($32 \times 95 = 3040$). Consequently, we must address an ill-posed problem.

All the approaches have been implemented with the same initialization: zero as mean and one as variance of the unknown $\mathbf{x}$, the hyperparameters σ_ϵ^2, σ_p^2 and ν set to 1, 0.05 and 0.1. We show in Fig. 2 the true image and reconstructions obtained by the five algorithms mentioned above.

The reconstruction results shown in Fig. 2(e) and Fig. 2(f) have the similar qualities to that obtained by the two classical variational Bayesian approaches, which are given in Fig. 2(c) and Fig. 2(d). The asymptotic results obtained by MCMC approach are theoretically the best ones since it leads to the true posterior distribution rather than its approximation. However in limited-time, the obtained samples are not able to fit its asymptotic results, which explains the relatively bad reconstruction shown in Fig. 2(b).

Table 1: PERFORMANCE COMPARISON IN TERMS OF PSNR(dB)/CPU TIME(S).

Method	MCMC	VBFS	VBPS	Grad	Proposed
PSNR(dB)	28.8	35.1	35.2	35.9	35.9
Time(s)	69313.8	723.5	327.6	23.4	3.2

We also provide numerical results indicating the advantage of our proposed algorithm in terms of computation time. The peak signal to noise ratios (PSNR) and the computation time are given in Tab. 1. We can see that our proposed approach has managed to achieve the reconstruction of highest PSNR (35.9dB) by taking just 3.2 seconds, which is 7 times quicker than the gradient-like algorithm, and more than 100 times faster than the classical VBAs and 20000 faster than the MCMC Gibbs approach for this small tomographic problem.

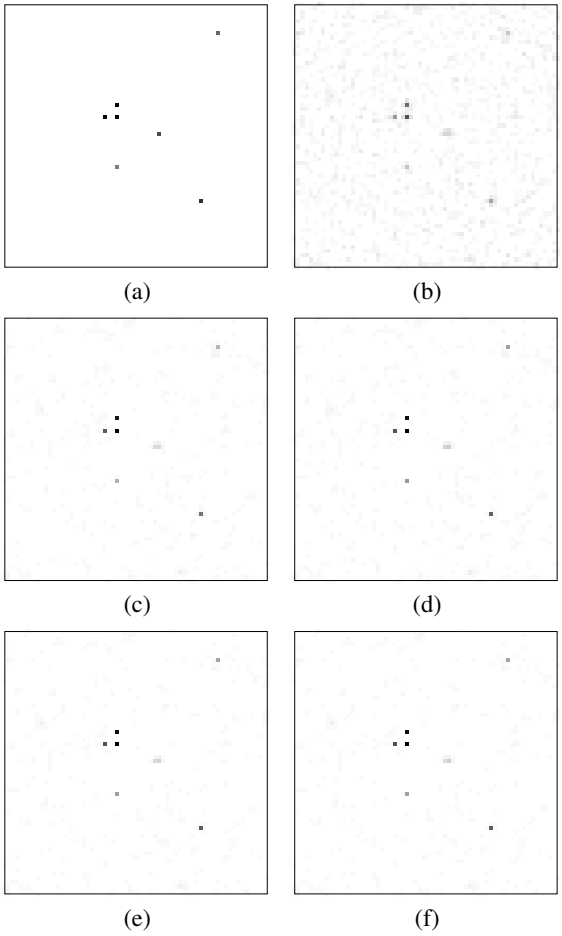

Fig. 2: Images are presented with the same inverse grayscale: (a) True image, (b) MCMC Gibbs approach, (c) classical VBA with hypothesis of full separability (16) (VBFS), (d) classical VBA with hypothesis of partial separability between $\mathbf{x}$ and $\mathbf{z}$ (VBPS), (e) gradient-like VBA (Grad) (f) our approach.

6. CONCLUSIONS

One efficient iterative variational Bayesian approach based on the subspace optimization method has been proposed. Numerical experimental results on a tomographic problem have been given to illustrate the advantage of our approach in terms of time efficiency. Since our proposed approach can be performed without any matrix inversion and in parallel for all the separated terms, it can be employed as well for large dimensional inverse problems.

7. REFERENCES

[1] C. P. Robert and G. Casella, *Monte-Carlo Statistical Methods*, Springer Texts in Statistics. Springer, New York, 2000.

[2] D. J. C. MacKay, "Ensemble learning and evidence maximization," Tech. Rep., Proc. NIPS, 1995.

[3] V. Šmídl and A. Quinn, *The Variational Bayes Method in Signal Processing*, Springer, 2006.

[4] A. Fraysse and T. Rodet, "A gradient-like variational bayesian algorithm," in *IEEE Workshop on Statistical Signal Processing (SSP)*, 2011, pp. 605–608.

[5] A. Fraysse and T. Rodet, "A measure-theoretic variational Bayesian algorithm for large dimensional problems.," Tech. Rep., 2012, http://hal.archives-ouvertes.fr/docs/00/70/22/59/PDF/var_bayV8.pdf.

[6] A. Miele and JW Cantrell, "Study on a memory gradient method for the minimization of functions," *Journal of Optimization Theory and Applications*, vol. 3, no. 6, pp. 459–470, 1969.

[7] G. Narkiss and M. Zibulevsky, *Sequential Subspace Optimization Method for Large-Scale Unconstrained Problems*, Technion-IIT, Department of Electrical Engineering, Oct. 2005.

[8] E. Chouzenoux, J. Idier, and S. Moussaoui, "A majorize-minimize strategy for subspace optimization applied to image restoration," *IEEE Transactions on Image Processing*, vol. 20, no. 18, pp. 1517–1528, 2011.

[9] R. A. Choudrey, *Variational Methods for Bayesian Independent Component Analysis*, Ph.D. thesis, University of Oxford, 2002.

[10] J. Kivinen and M. Warmuth, "Exponentiated gradient versus gradient descent for linear predictors," *Information and Computation*, vol. 132, no. 1, pp. 1–63, 1997.

[11] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders, "Variational Bayesian image restoration based on a product of t-distributions image prior," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1795–1805, Oct. 2008.

[12] S. Jana and P. Moulin, "Optimality of KLT for high-rate transform coding of gaussian vector-scale mixtures: Application to reconstruction, estimation, and classification," *IEEE Transactions on Information Theory*, vol. 52, no. 9, pp. 4049–4067, 2006.

[13] M. J. Wainwright and E. P. Simoncelli, "Scale mixtures of gaussians and the statistics of natural images," *Advances in neural information processing systems*, vol. 12, no. 1, pp. 855–861, 2000.