A gradient-like variational Bayesian algorithm

Aurélia Fraysse, Thomas Rodet

A GRADIENT-LIKE VARIATIONAL BAYESIAN ALGORITHM

Aurélia Fraysse, Thomas Rodet

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

ABSTRACT

In this paper we provide a new algorithm allowing to solve a variational Bayesian issue which can be seen as a functional optimization problem. The main contribution of this paper is to transpose a classical iterative algorithm of optimization in the metric space of probability densities involved in the Bayesian methodology. Another important part is the application of our algorithm to a class of linear inverse problems where estimated quantities are assumed to be sparse. Finally, we compare performances of our method with classical ones on a tomographic problem. Preliminary results on a small dimensional example show that our new algorithm is faster than the classical approaches for the same quality of reconstruction.

Index Terms— Variational Bayesian, infinite dimensional optimization, sparse reconstruction.

1. INTRODUCTION

The recent advances of information technologies have widely increased the size of data involved in reconstruction problems. Simultaneously, signal processing techniques have allowed to overcome instrumentation limitations, creating hence new theoretical challenges. In particular, the size of datasets collected nowadays can be very large. There is therefore a need for reconstruction methods for large dimensional inverse problems.

A classical approach when dealing with these ill posed problems is to introduce additional information. The Bayesian methodology involved in this paper consists in a modelisation of sources as probability density functions, see for instance [4] for details. This approach allows the development of unsupervised methods, where the so called hyperparameters, i.e. parameters of the model, are adjusted automatically to tune the weight between the *a priori* information and the information coming from the data. We call these methods "fully Bayesian" as they consist in a construction of a posterior distribution of parameters of interest and of hyperparameters. In practice, most of "fully Bayesian" approaches use Markov Chain Monte Carlo (MCMC) [9] algorithms to estimate the posterior mean. For instance, in the case of deconvolution problems, where the covariance matrix is easily invertible, efficient samplers can be developed and this method can easily be handled. So there are many MCMC approaches that developed fully Bayesian [5]. However, in general, the use of MCMC is limited by the lack of an effective sample of correlated vectors.

Therefore D. MacKay proposed in 1995 an alternative methodology, the so called variational Bayesian method [7]. The main idea of this method is to approximate the posterior distribution by a separable density. Even if it gives approximate solutions, this method could be more efficient that MCMC in large dimensional cases, especially when the covariance matrix is no longer invertible. Indeed, as the calculations are analytical, the rate of convergence is much better than for the MCMC approaches. This methodology is applied in lot of areas: sources separations using ICA [3], deconvolution [2], recursive methods [11]. However, as we will see later, variational Bayesian method leads to an implicit solution. Hence iterative methods are used to approximate this solution. And classical iterative methods used in this context are often too heavy to be efficient for large dimensional dataset.

The main contribution of this paper is to define an iterative algorithm able to provide, in few iterations, a close approximation of the solution of the variational Bayesian problem. The original idea is to adapt a classical finite-dimensional optimization algorithm, the gradient descent method [8], to the space of probability distributions. Another contribution of this paper is the application of our method to a class of linear inversion problems involving sparse prior information. As an example, we apply this new algorithm to a classical problem of tomography.

In section 2 we recall the classical variational Bayesian approach whereas in section 3 we introduce our algorithm. Section 4 gives an application of this method on a linear problem with sparse information illustrated in section 5. Finally, Section 6 concludes the paper.

2. BAYESIAN VARIATIONAL METHODS

We first introduce the key principle of the variational Bayesian method presented in [3]. This Bayesian method is mainly used for ill-posed inverse problems where the posterior distribution takes intricate forms. The main idea is to approximate the true posterior by separable distributions, close to this posterior in the sense of the Kullback-Leibler divergence. This approximation step turns the estimation problem as an optimization paradigm which enlarges the range of validity of Bayesian methods in terms of complexity of the inverse problem.

In the following we denote by $\mathbf{Y} \in \mathbb{R}^M$ the M dimensional vector containing the data information whereas $\mathbf{W} \in \mathbb{R}^N$ represents the vector of hidden variables to be determined. We assume that $\mathbf{W}$ is random with a known distribution $p(\mathbf{W})$, the *a priori* distribution. The main challenge is to determine the corresponding posterior distribution $p(\mathbf{W}|\mathbf{Y})$. Note that even for a simple *a priori* distribution, the posterior distribution can have an intricate form. We thus have to approximate it by a separable probability density q.

To determine this approximating law, we first consider the log-likelihood of data which can be written, see [3], as:

$$\log p(\mathbf{Y}) = F(q(\mathbf{W})) + \mathcal{KL}[q(\mathbf{W})||p(\mathbf{W}|\mathbf{Y}))], \qquad (1)$$

where $\mathcal{KL}[q(\mathbf{W})||p(\mathbf{W}|\mathbf{Y}))$ is the Kullback-Leibler divergence between the approximate probability density function (pdf) q and the

posterior pdf. In this case,

$$F(q(\mathbf{W})) = \int q(\mathbf{W}) \log \left(\frac{p(\mathbf{Y}, \mathbf{W})}{q(\mathbf{W})} \right) d\mathbf{W}, \tag{2}$$

is the "negative free energy". As $\log p(\mathbf{Y})$ is independent of the approximating density q, minimizing the Kullback-Leibler divergence is obviously equivalent to maximize this negative free energy.

The objective of variational Bayesian methods is thus to find

$$q^{opt} = \arg\max_{q} F(q(\boldsymbol{W})). \tag{3}$$

The negative free energy can also be written as

$$F(q(\mathbf{W})) = \langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{q(\mathbf{W})} + \mathcal{H}(\mathbf{W}), \tag{4}$$

where $\mathcal{H}(\mathbf{W})$ is the entropy of $\mathbf{W}$ under the distribution q, whereas

$$\langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{q(\mathbf{W})} = \int \log(p(\mathbf{Y}, \mathbf{W})) q(\mathbf{W}) d\mathbf{W}. \tag{5}$$

represent the expectation of $\log p(\mathbf{Y}, \mathbf{W})$ under the distribution $q(\mathbf{W})$. Note that as $\mathcal{K}L$ is convex, Eq. (1) ensures that F is concave relatively to the approximating probability density function $q(\mathbf{W})$, thus we have to solve a convex infinite dimensional optimization problem.

Assuming that q is a separable pdf, i.e. $q(\mathbf{W}) = \prod_i q_i(w_i)$, we can obtain an analytic form for $q_i(w_i)$ (see [3] for details on the variational calculus):

$$q_i(w_i) = \frac{1}{K_i} \exp \left(\langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{\prod_{j \neq i} q_j(w_j)} \right). \tag{6}$$

Although this solution is obtained analytically, Eq. (6) clearly does not have an explicit form. This solution is hardly tractable in practice, and is thus approximated thanks to iterative methods. These methods impose the use of conjugate prior to obtain a posterior law belonging to a known family. In this context, optimizing the posterior turns out to an optimization of its distribution parameters. As in Eq. (6) the calculus of q_i imposes the knowledge of all q_j for j different from i, this optimization is either performed alternatively or by groups of coordinates, by storing the corresponding covariance matrix.

However, this method increases considerably the computation time. To reduce this drawback, we can only perform the optimization algorithm by group of coordinates. This approach reduces the number of iterations but induces to store and invert a large correlation matrix. Hence for large dimensional problems these methods are not efficient in practice. Our purpose is thus to solve the functional optimization problem given by the Bayesian variational method more efficiently than the approaches induced by Eq. (6).

3. THE PROPOSED ALGORITHM

The optimization problem involved in variational Bayesian method is an infinite dimensional concave problem. It would therefore be convenient to determine the approximating density thanks to classical optimization algorithms, such as the gradient descent method. This is this method which is employed hereafter. However, we have to pay a particular attention to the fact that we stand in an infinite dimensional non-vector space: the space of probability density functions. There are two ways to understand this issue. The first one is to consider that we treat a subspace of the L^1 function space which is an infinite dimensional vector space. In this case the classical gradient descent method is still feasible. However, we thus have to pay a particular attention to the fact that all elements of this subspace have to satisfy $\int f = 1$, which induces a projection step at each iteration. The second approach, developed here, is to consider that we stand in a subspace of the probability measures space. The main advantage is that the normalization step is no longer necessary. The main drawback is that a measure space is no longer a vector space (see [1] for details). We thus have to adapt the gradient descent method in this case, taking the structure of the space into account.

Let us define the proposed method. Assume that for $k \geq 0$, $\{q_1^k, \ldots, q_N^k\}$ are constructed and that $q^k(\mathbf{W}) = \prod q_i^k(w_i)$. As we stand in the space of probability measures, the following step must give a probability density on $\mathbb{R}^N$, absolutely continuous with respect to $q^k d\lambda$, λ being the Lebesgue measure on $\mathbb{R}^N$. Such a condition is satisfied, thanks to the Radon-Nikodym theorem, see [10] for instance, if we consider

$$q^{k+1} = h q^k \tag{7}$$

where $h \in L^1(q^k)$ is a positive separable function. As in the gradient descent algorithm, this function h is based on the Gateaux derivatives of F at q^k. In order to ensure that h is a positive integrable function we choose to take

$$h(\mathbf{W}) = \exp(\alpha \nabla F(q(\mathbf{W}))). \tag{8}$$

where ∇F stands for the Gateaux derivative of F whereas $\alpha > 0$ is the algorithm step-size. We take this form for h and we choose α small enough to ensure that the functional F increases at each iteration. Furthermore a calculus similar to those of [3] shows that

$$\forall i = 1, \ldots, N \quad \frac{\partial F}{\partial q_i} = \langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{\prod_{j \neq i} q_j^k(w_j)} - \log q_i - 1.$$

This entails

$$\begin{aligned} h_i(w_i) &= \left(\frac{1}{K_i} \frac{\exp \left(\langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{\prod_{j \neq i} q_j^k(w_j)} \right)}{q_i^k} \right)^{\alpha} \\ &= \left(\frac{q_i^r(w_i)}{q_i^k(w_i)} \right)^{\alpha}, \end{aligned} \tag{9}$$

where $q_i^r(w_i) = \frac{1}{K_i} \exp \left(\langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{\prod_{j \neq i} q_j^k(w_j)} \right)$ is an intermediate density measure.

We thus define q^{k+1} as

$$\begin{aligned} q^{k+1}(\mathbf{W}) &= q^k(\mathbf{W}) \left(\prod_i \frac{1}{K_i} \frac{\exp \left(\langle \log p(\mathbf{Y}, \mathbf{W}) \rangle_{\prod_{j \neq i} q_j^k(w_j)} \right)}{q_i^k} \right)^{\alpha} \\ &= q^k(\mathbf{W}) \left(\frac{q^r(\mathbf{W})}{q^k(\mathbf{W})} \right)^{\alpha}. \end{aligned} \tag{10}$$

This algorithm allows to minimize jointly all (q_i) unlike the classical Bayesian Variational algorithm. Moreover, the stepsize α can be chosen in order to optimize the convergence rate. Note that with a logarithmic scale we retrieve the classical updating equation of the gradient descent method.

4. APPLICATION TO SPARSE LINEAR PROBLEMS

In order to have a better understanding of the algorithm defined in section 3, we show how it can be applied to linear inverse problems.

4.1. The model

We treat in this section the classical linear problem:

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{b}, \tag{11}$$

where $\mathbf{H}$ is a matrix in $\mathcal{M}_{N\times M}$ whereas $\mathbf{b} \in \mathbb{R}^M$ is a Gaussian white noise. Here the parameter vector $\mathbf{X}$ is assumed to be separable. Concerning the prior distribution we choose to take sparsity into account by considering that the distribution of $\mathbf{X}$ is a separable Student-t distribution. Indeed, Student-t distributions is a large class of distributions depending on a parameter. For small values of this parameter, they are heavy-tailed distributions, see for instance [6] for details. In the following, we use the fact that a Student-t distribution can be modelised as a Gaussian Scale Mixture, that is a Gaussian distribution with an inverse variance given by a hidden variable following a Gamma law, $\mathcal{G}amma(\frac{a}{2}, \frac{a}{2})$. Thus, for every $i = 1, \ldots, N$, we take

$$p(x_i) = \int_{\mathbb{R}^+} \frac{\sqrt{z_i}}{(2\pi)^{N/2}|\sigma_1^2|^{1/2}} e^{-\frac{z_i x_i^2}{2\sigma_1^2}} \frac{(\frac{a}{2})^{a/2} z_i^{a/2-1} e^{-\frac{az_i}{2}}}{\Gamma(\frac{a}{2})} dz_i.$$

Hence, we choose to solve an extended problem which takes the hidden vector $\mathbf{Z}$ into account. Thanks to this rewriting, the Student-t distribution is conjugate with the Gaussian likelihood.

In this setting, one can easily check that the joint posterior distribution is given by

$$\begin{aligned} p(\mathbf{x}, \mathbf{z}|\mathbf{y}) &\propto \exp\left[-\frac{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{x}\|^2}{2\sigma_b^2}\right] \\ &* \prod_i \frac{\sqrt{z_i}}{\sigma_1} \exp\left[-\frac{z_i x_i^2}{2\sigma_1^2}\right] \frac{(\frac{a}{2})^{a/2} z_i^{a/2-1} e^{-\frac{az_i}{2}}}{\Gamma(\frac{a}{2})}. \end{aligned} \tag{12}$$

This posterior distribution is not tractable analytically due to two main drawbacks. The first issue is the link between $\boldsymbol{X}$ and $\boldsymbol{Z}$, which is solved by the classical variational Bayesian approach. The second one occurs when the dimension of the vector $\boldsymbol{X}$ increases. In this case, the correlation matrix is too large to be inverted efficiently. This issue is solved by our algorithm presented in section 3. Details are exposed hereafter.

4.2. Variational Bayesian algorithm

In this context we apply the algorithm introduced earlier, by taking $\mathbf{W} = (\mathbf{X}, \mathbf{Z})$. We want to approximate (12) by separable laws, thus by a probability distribution

$$q(\mathbf{W}) = q(\mathbf{X}, \mathbf{Z}) = \prod_i q_i(x_i)\tilde{q}_i(z_i),$$

which maximizes (4).

As we can see in (12), for $i = 1, \ldots, N$, the posterior law of X_i is Gaussian whereas the posterior law of Z_i is Gamma. Therefore we determine $\boldsymbol{X}$ thanks to our method and we update afterward the parameters of $\boldsymbol{Z}$. We choose to initialize our approximating laws by taking $q_i^0(x_i)$ as a Gaussian probability density function and the approximate law $\tilde{q}_i^0(z_i)$ as a Gamma one. Thus, for $i = 1, \ldots, N$ we take:

$$\begin{aligned} q_i^0(x_i) &= \mathcal{N}(m_0(i), \sigma_0^2(i)) \\ \tilde{q}_i^0(z_i) &= \mathcal{G}amma(a_0, b_0). \end{aligned}$$

As mentioned in part 2, from the conjugate hypothesis, at each iteration, q_i^k stays a Gaussian distribution whereas $\tilde{q}_i^k$ stays a Gamma law. At the next step, the density of q_i^{k+1} which depends of the step size α, is computed with Eq. (10). We see that, for $i = 1, \ldots, N$,

$$q_i^{k+1}(\alpha) = q_i^k \left(\frac{q_i^r}{q_i^k}\right)^\alpha$$

is still a Gaussian law with variance:

$$\sigma_{k+1}^2(i) = \frac{\sigma_r^2(i)\sigma_k^2(i)}{\sigma_r^2(i) + \alpha(\sigma_k^2(i) - \sigma_r^2(i))}$$

where

$$\sigma_r^2(i) = \left(\frac{(\boldsymbol{H}^T\boldsymbol{H})[i,i]}{\sigma_b^2} + \frac{a_k(i)}{b_k(i)\sigma_1^2}\right)^{-1} \tag{13}$$

and the mean of $q_i^{k+1}(\alpha)$ is

$$m_{k+1}(i) = \frac{m_k(i)\sigma_r^2(i) + \alpha(m_r(i)\sigma_k^2(i) - m_k(i)\sigma_r^2(i))}{\sigma_r^2(i) + \alpha(\sigma_k^2(i) - \sigma_r^2(i))},$$

with

$$m_r(i) = \sigma_r^2(i) \times \left(\frac{\boldsymbol{H}^T\boldsymbol{y} - (\boldsymbol{H}^T\boldsymbol{H} - \mathrm{diag}(\boldsymbol{H}^T\boldsymbol{H}))\boldsymbol{m}_k}{\sigma_b^2}\right)_i \tag{14}$$

Performances of this algorithm strongly depend on the step size α. For a fixed α small enough, this algorithm indeed converges. However, in order to increase the speed of convergence, we choose to determine an approximation of the optimal step size α_{opt} thanks to a Taylor expansion of our functional. Finally for every $i = 1, \ldots, N$, we take $q_i^{k+1} = q_i^{k+1}(\alpha_{opt})$.

Concerning the approximation of $\tilde{q}_i$ we keep the standard variational Bayesian method. We obtain a Gamma function with updating equations:

$$a_{k+1}(i) = \frac{a}{2} + \frac{1}{2}, \tag{15}$$

$$b_{k+1}(i) = \frac{m_k^2(i) + \sigma_k^2(i)}{2\sigma_1^2} + \frac{a}{2}. \tag{16}$$

5. RESULTS

5.1. Simulation parameters

In this section we emphasize our approach by comparing it with classical Bayesian methods, i.e. MCMC approach and classical variational Bayesian (VB), and with a classical non bayesian reconstruction method, the filtered Back Projection (FBP method). We choose to treat the linear problem given by Eq. (11) with a non invertible matrix $\boldsymbol{H}$ coming from a tomographic problem. From the limitations of MCMC approach, we solve a relatively small inverse problem (image $64 \times 64 = 4096$ unknowns).

The test image is given by a sparse phantom, composed of 7 peaks with a magnitude between 0.5 and 1 (see Fig.2(a)). We have simulated data in parallel beam geometry. These projections are collected from 32 angles, uniformly spaced over $[0, 180[$. Each projection is composed of 95 detector cells. We add a white Gaussian noise (iid) with standard deviation equal to 0.3 (see Fig. 1). Data have thus a relatively bad signal to noise ratio and the number of unknowns is larger than the number of data, which leads to an ill-posed inverse problem.

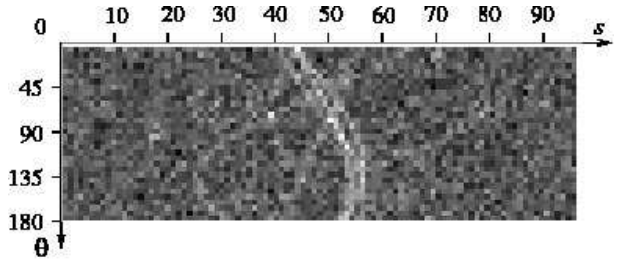

Fig. 1. Data collected : sinogram composed of 32 angles and 95 detector cells.

5.2. Results and discussion

All the iterative approaches are initialized with a zero mean and a variance equal to one, and the hyper-parameters σ_b^2, σ_1^2 and a are respectively fixed to 1, 0.05 and 0.1. The original image and its different reconstructions are summed up on Fig. 2. A comparison of Fig. 2 (b) with 2 (c), 2 (d) and 2 (e) clearly shows that the analytical inversion of the Radon transform perform using the Filtered Back Projection (FBP) algorithm is less robust to noise than Bayesian approaches. Asymptotically, in Bayesian cases theoretical results are favorable to the MCMC approach, as it does not need any approximation. In practice, the number of samples is too small to fit with asymptotic results of MCMC method, which explains the bad reconstruction observed in Fig. 2(c). Finally, our approach (see Fig. 2(e)) has the same reconstruction quality than the classical variational Bayesian approach (see Fig. 2(d)). However when we compare the execution time (see Tab. 1), we see that our approach is 5 time faster than the VB approach and 370 faster than the MCMC approach for this small inverse problem. Moreover this ratio increases with the size of the problem as both MCMC and classical variational Bayesian need the inversion of a covariance matrix at each iteration. It is not the case of our algorithm. Thanks to this benefit, large dimensional problems can be solved by our fully Bayesian approach.

Table 1. Computing time (s).

method	FBP	VB	our approach	MCMC Gibbs
CPU time (s)	0.05	586.20	103.55	37079.50

6. CONCLUSION

In this paper, we have defined a new iterative algorithm based on the descent gradient principle in the space of probability densities. We have also shown how this algorithm can be implemented in the context of variational Bayesian methods. The main interest of this algorithm is that it converges faster than the classical Bayesian methods and allows an use on large dimensional datasets. A small tomographic application allows us to compare our method with classical ones. We see that even in small cases, performances of our algorithm can be better than classical ones. Furthermore its linear structure simplifies an use on large dimensional problems.

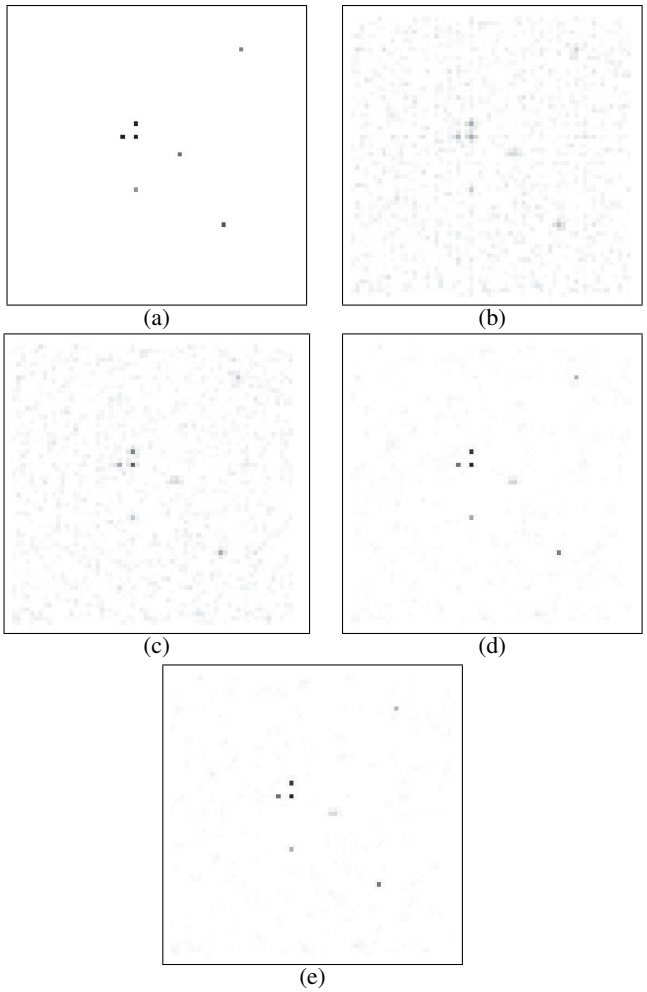

Fig. 2. Images are presented with the same inversed grayscale: (a) true image of 7 peaks, (b) FBP with ramp filter, (c) MCMC Gibbs approach, (d) classical variational Bayesian, (e) our method.

7. REFERENCES

[1] P. Billingsley. *Convergence of probability measures*. Wiley edition, 1999.

[2] G. Chantas, N. Galatsanos, A. Likas, and M. Saunders. Variational Bayesian image restoration based on a product of t-distributions image prior. *IEEE Trans. Image Processing*, 17(10):1795–1805, October 2008.

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

[4] G. Demoment. Image reconstruction and restoration: Overview of common estimation structure and problems. *IEEE Trans. Acoust. Speech, Signal Processing*, ASSP-37(12):2024–2036, December 1989.

[5] J.-F. Giovannelli. Unsupervised bayesian convex deconvolution based on a field with an explicit partition function. *IEEE Trans. Image Processing*, 17(1):16–26, January 2008.

[6] E. Grosswald. The student t-distribution of any degree of freedom is infinitely divisible. *Proba. Theory and Rel. Fields*, 1976.

[7] D. J. C. MacKay. Ensemble learning and evidence maximization. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.54.4083, 1995.

[8] J. Nocedal and S. J. Wright. *Numerical Optimization*. Series in Operations Research. Springer Verlag, New York, 2000.

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

[10] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, 1987.

[11] V. Smidl and A. Quinn. Variational bayesian filtering. *IEEE Trans. Signal Processing*, 56(10):5020–5030, Oct. 2008.