A FAST VARIATIONAL APPROACH FOR BAYESIAN COMPRESSIVE SENSING WITH INFORMATIVE PRIORS

Evripidis Karseras and Wei Dai

Department of Electrical and Electronic Engineering
Imperial College, London, UK
{e.karseras11, wei.dai1}@imperial.ac.uk

ABSTRACT

The Sparse Bayesian learning (SBL) framework has been successfully adopted for sparse signal recovery. In SBL inference can be performed either via *Type-II Maximum Likelihood* or by following a *Variational approach*. When employing *uninformative* prior distributions, fast algorithms have been proposed for both renditions of SBL and it has been proven that they are equivalent. Unfortunately the use of such priors prohibits the incorporation of prior statistical information which can be beneficial in terms of convergence and accuracy. A modified variational approach is proposed, resulting in a fast variational algorithm for *informative* priors. A fixed point analysis is performed with the major challenge being the highly involved analytical expressions for the points in the fixed set. The given theoretical analysis demonstrates how this issue can be circumvented. Comprehensive empirical results are given to support the claims.

Index Terms— sparse Bayesian learning, variational RVM, fast RVM, informative priors

1. INTRODUCTION

Sparse Bayesian Learning (SBL) was introduced in [1] and accomplishes via a hierarchy of distributions to produce highly sparse models for the input. An inference algorithm is derived to recover the most probable values for the model parameters and the controlling hyper-parameters. This is best known as *Type-II Maximum Likelihood*. Under the assumption that the hyper-prior distributions are *uninformative* a fast version of this algorithm was proposed in [2] and [3]. Its use for sparse signal recovery was showcased in [4].

A fully Bayesian treatment was introduced in [5] with the variational rendition of SBL (VSBL). The VSBL provides estimates for the distributions of both the model parameters and *hyper-parameters* as opposed to *Type-II ML* which assumes point estimates for the hyper-parameters. This extension makes SBL more flexible and the recovery algorithm more controllable; something intractable for *Type-II ML*. Under the *uninformative* assumption a fast VSBL algorithm was proposed in [6]. It was proven that the fast algorithms for VSBL and *Type-II ML* are equivalent.

We are concerned with cases where previous knowledge for a sparse signal needs to be incorporated into the recovery algorithm. Prior information can result in improved convergence speed and reconstruction accuracy. The VSBL approach is fitting for these scenarios but regrettably the fast algorithms can no longer be used. The method by which complexity is reduced in the fast VSBL algorithm unfortunately cannot be employed because of the bias introduced by the *informative* prior.

In this paper we perform a fixed point analysis of VSBL as a first step to control complexity. This was primarily the case in [6] for the *uninformative* case. The major challenges in our approach are the highly complicated mathematical expressions for when an *informative* prior is used. These prohibited the derivation of any results towards speeding up inference. The given theoretical study proposes advantageous workarounds that are practical and computationally efficient. We demonstrate that fast VSBL is in fact possible for *informative* priors and that complexity can be controlled.

In Section 2 a summary of the fast VSBL is given. Follows in Section 3, a rigorous analysis of fast VSBL for informative priors. We provide theoretical results that support a fast approach. In Section 4 ample empirical evidence on synthetic data is provided to support our claims. The performance of the modifications is compared against the conventional VSBL in terms of the strength of the prior, convergence speed, reconstruction error and problem size.

2. VARIATIONAL BAYESIAN LEARNING

The learning process is concerned with recovering the most probable values for the model parameters $\boldsymbol{x}$ for a specific input $\boldsymbol{y}$. The investigated models deal with cases where the parameter vector is *sparse*. In mathematical terms,

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x} + \boldsymbol{n} \tag{1}$$

where $\boldsymbol{\Phi} \in \mathbb{R}^{M \times N}$ and $\boldsymbol{n} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$.

978-1-4799-2893-4/14/$31.00 ©2014 Crown

2.1. Variational Inference

In SBL [1] one separate hyper-parameter α_i is contemplated to control the variance of each component x_i:

$$p\left(\boldsymbol{x}|\boldsymbol{\alpha}\right)=\prod_{i=1}^{n}\mathcal{N}\left(x_i|0,\alpha_i^{-1}\right)=\mathcal{N}\left(\boldsymbol{x}|\boldsymbol{0},\boldsymbol{A}^{-1}\right)$$

where matrix $\boldsymbol{A}=\mathrm{diag}\left([\alpha_1,\cdots,\alpha_n]\right)$. Hyper-prior distributions are also defined as $p(\alpha_i)=\mathrm{Gamma}(\alpha_i|a,b)=b^a\alpha_i^{a-1}e^{-b\alpha}/\Gamma(a)$ where $\Gamma(a)$ is the Gamma function [1].

In the uninformative case $a=b=0$ expresses no prior biases, i.e. *uninformative*. From the Bayes rule the posterior is $p(\boldsymbol{x},\boldsymbol{\alpha},\sigma^2|\boldsymbol{y})=p(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\alpha},\sigma^2)p(\boldsymbol{\alpha},\sigma^2|\boldsymbol{y})$. While the first part of the right-hand side is tractable, an approximation to the second part has to be found by the maximum likelihood solution of $p\left(\boldsymbol{y}|\boldsymbol{\alpha},\sigma^2\right)=\int p(\boldsymbol{y}|\boldsymbol{x},\sigma^2)p(\boldsymbol{x}|\boldsymbol{\alpha})d\boldsymbol{x}$. This is widely known as *Type-II Maximum Likelihood*. We disregard modelling the noise without compromising generality.

In the informative case, i.e., VSBL [5], it is assumed that $a_i\geq 0$ and $b_i\geq 0$. An approximation to the posterior $p(\boldsymbol{x},\boldsymbol{\alpha},\sigma^2|\boldsymbol{y})\approx p(\boldsymbol{x},\boldsymbol{\alpha},\sigma^2)=q(\boldsymbol{x})q(\boldsymbol{\alpha})q(\sigma^2)$ is used with $q(\boldsymbol{x})=\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\boldsymbol{\Sigma})$, $q(\boldsymbol{\alpha})=\prod_{i=1}^{N}\mathrm{Ga}(\alpha_i|\tilde{a}_i,\tilde{b}_i)$ where the parameters are given by the well-known update formulae:

$$\boldsymbol{\Sigma}=\left(\sigma^2\boldsymbol{\Phi}^T\boldsymbol{\Phi}+\tilde{A}\right)^{-1},\quad \boldsymbol{\mu}=\sigma^2\boldsymbol{\Sigma}\boldsymbol{\Phi}^T\boldsymbol{y} \tag{2}$$

$$\tilde{a}_i=a_i+\frac{1}{2},\ \ \tilde{b}_i=b_i+\frac{|x_i|^2+\boldsymbol{\Sigma}_{ii}}{2},\ \ \tilde{\alpha}_i=\frac{\tilde{a}_i}{\tilde{b}_i}. \tag{3}$$

We redirect to [5] for the noise distribution parameter update expressions. The expression for $\tilde{\alpha}_i$ is the mean $\mathbb{E}[\alpha_i]$ of a Gamma distribution.

In this sense the VSBL is more general, acting in a fully Bayesian framework by providing closed form solutions to approximations of distributions that are intractable to derive analytically. The VSBL is ideal for when the estimates of the hyper-prior distributions are required, something which becomes unmanageable within the *Type-II ML* framework. The price to pay is higher computational complexity. This case will occupy us for the rest of the paper.

2.2. The Fast Variational Approach

By maintaining $a_i=b_i=0,\ \forall\ i\in[1,N]$ the authors in [6] manage to decouple the estimated hyper-parameters from each other in a way similar to [3] and control complexity. The non-linear map for a specific hyper-parameter at iteration m is derived,

$$\begin{aligned}\frac{1}{\tilde{\alpha}_i^{[m+1]}}&=\left[w_i^2+z_i-\frac{z_i^2+2z_iw_i^2}{\frac{1}{\tilde{\alpha}_i^{[m]}}+z_i}+\frac{z_i^2w_i^2}{(\frac{1}{\tilde{\alpha}_i^{[m]}}+z_i)^2}\right]\\&=F(\tilde{\alpha}_i^{[m]})\end{aligned} \tag{4}$$

where $z_i=\boldsymbol{e}_i^T\boldsymbol{\Sigma}_{-i}\boldsymbol{e}_i$, $w_i^2=\sigma^2\boldsymbol{e}_i^T\boldsymbol{\Sigma}_{-i}\boldsymbol{\Phi}^T\boldsymbol{y}\boldsymbol{y}^T\boldsymbol{\Phi}\boldsymbol{\Sigma}_{-i}\boldsymbol{e}_i$, $\boldsymbol{\Sigma}_{-i}=\left(\sigma^2\boldsymbol{\Phi}_{-i}^T\boldsymbol{\Phi}_{-i}+\tilde{\boldsymbol{A}}_{-i}\right)^{-1}$, $\tilde{\boldsymbol{A}}=\mathrm{diag}\left([\tilde{\alpha}_1,\cdots,\tilde{\alpha}_n]\right)$ and F is the map function. Notation $\boldsymbol{\Phi}_{-i}$ means the removal of column i while $\tilde{\boldsymbol{A}}_{-i}$ the removal of row and column i. Fixed point analysis is performed by letting $m\to\infty$. At convergence it holds that $\tilde{\alpha}_i^{[m+1]}=\tilde{\alpha}_i^{[m]}=\tilde{\alpha}_i^*$. By solving $\tilde{\alpha}_i^*-F\left(\tilde{\alpha}_i^*\right)=0$ two fixed points are found which are asymptotically stable:

$$\tilde{\alpha}_i=\begin{cases}(w_i^2-z_i)^{-1}, & w_i^2>z_i\\ +\infty & w_i^2\leq z_i.\end{cases} \tag{5}$$

It was proven in [6] that the fast VSBL algorithm is equivalent to the fast *Type-II ML* algorithm in [3]. When $\tilde{\alpha}_i^{-1}=0$ then it is implied that $x_i=0$ and the derivation of a closed form expression for the fixed point makes it possible to reduce complexity in the VSBL by avoiding extraneous iterations.

3. FAST VSBL FOR INFORMATIVE PRIORS

The uninformative assumption is proven to result in efficient algorithms for sparse signal recovery. Problems arise with the need to incorporate prior knowledge into the model. This is particularly the case when recovering time series of sparse signals where past information can be used to improve the speed and the accuracy of convergence.

An *informative* prior can control the uncertainty but this forces us to abandon the fast algorithm for variational inference [6]. This is verified by inspecting Equations (3). Assuming that[1] $a_i>0, b_i>0$ the value of $\tilde{\alpha}_i^{-1}=0$ becomes unattainable making the pruning rule in Equation (5) obsolete. Computational complexity then increases and becomes worse than iteratively updating Equations (2) and (3).

3.1. Introducing prior knowledge

We follow [6] by adopting a fixed point analysis. Regrettably, when $a_i>0, b_i>0$ the analysis complicates considerably since the bias introduced by the prior results in the expression given below,

$$\begin{aligned}\frac{(2a_i+1)}{\tilde{\alpha}_i^{[m+1]}}&=\left[2b_i+w_i^2+z_i-\frac{z_i^2+2w_i^2z_i}{\frac{1}{\tilde{\alpha}_i^{[m]}}+z_i}+\frac{z_i^2w_i^2}{(\frac{1}{\tilde{\alpha}_i^{[m]}}+z_i)^2}\right]\\&=G(\tilde{\alpha}_i^{[m]}).\end{aligned} \tag{6}$$

To derive the fixed points of the above map function we attempt solving the following,

$$\tilde{\alpha}_i^*-G\left(\tilde{\alpha}_i^*\right)=0. \tag{7}$$

We discover that the choice of an *uninformative prior* resulted in the *well-posed* polynomial of Equation (4) which could be

[1] Only the range of values where $a_i>0, b_i>0$ are considered for which the Gamma distribution is well defined.

rapidly factorised. For the *informative* case this convenience vanishes. Attempts for analysing the closed form solutions can be found in [7, Appx. B] but the results are not of practical or theoretical use. Nevertheless, the following theorem substantially simplifies the computation of the fixed points.

Theorem 1. *Assume a Gamma hyper-prior for α_i with $a_i > 0, b_i > 0$ and that the values of $\alpha_{j\neq i}$ are fixed. Let $\tilde{\alpha}_i^*$ be the value at which the map function G converges when iteration $m \to \infty$. Then $\beta_i^* = \frac{1}{\tilde{\alpha}_i^*} + z_i$ is one of the three solutions of the cubic polynomial*[2]*,*

$$f(\beta_i^*) + g(\beta_i^*) = 0 \tag{8}$$

that satisfies $\beta_i^* > z_i$[3]*, where $f(\beta_i^*) = \left(\beta_i^* - w_i^2\right)\left(\beta_i^* - z_i\right)^2$ and $g(\beta_i^*) = 2(\beta_i^*)^2\left(a_i(\beta_i^* - z_i) - b_i\right)$.*

Proof (sketch): Starting from Equations (3) the expression for $\tilde{\alpha}_i$ is formed for $a_i > 0, b_i > 0$ which is given in Equation (6). By applying the variable change and after some mathematical manipulation one arrives at the stated claim.

Theorem 1 provides a workaround for the intricacy of the analytical expressions. It is proposed that the solutions of Equation (7) can equivalently be acquired by solving Equation (8). We observe that Equation (7) can be divided into two distinct functions f, g; where the contribution from the prior is given by function $g(\beta_i^*)$. By setting $a_i = b_i = 0$ we fall back to solving $f(\beta_i^*) = 0$ which gives the fixed points in Equation (5).

3.1.1. Cardinality of the fixed points

Theorem 1 provides useful intuition for the fixed points and facilitates further qualitative analysis by using functions f and g. A full scale analysis is not collocated here due to space requirements. The following proposition has been proven.

Proposition 1. *Assume that the same conditions hold as in Theorem 1. Solving Equation* (7) *can result in one of the following cases:*

1. *if $w_i^2 \leq z_i$, then there exists* only one *valid root,*
2. *if $w_i^2 > z_i$, then for $\frac{b_i}{a_i} \geq \frac{2}{3}(w_i^2 - z_i)$ there exists* only one *valid root,*
3. *if $w_i^2 > z_i$, then for $\frac{b_i}{a_i} < \frac{2}{3}(w_i^2 - z_i)$ there* may *exist three distinct valid roots.*

By Proposition 1 it is possible for the fixed set to contain three *distinct real fixed points*. This case was a rarity during empirical tests. A numerical example can be constructed by setting $w_i^2 = 0.6, z_i = 0.4, a_i = 0.02$ and $b_i = 0.002$.

[2]Closed form solutions for the cubic equation can be found in algebra textbooks or on-line.

[3]Henceforth, we consider roots satisfying $\beta_i^* > z_i$ as valid roots.

3.1.2. Fixed point selection

Focusing on the third case where x_i is deemed to be non-zero but the exact value of $\tilde{\alpha}_i^*$ can be any of the three in the fixed set. In order to resort to a choice the stability of each fixed point has to be assessed. This requires the analytical expression of $\frac{dG}{d\tilde{\alpha}_i}$ to be derived for each of the fixed points. The highly complicated expressions unfortunately prohibit this analysis.

A reasonable choice is to choose the fixed point that maximises the variational lower bound,

$$\begin{aligned}\mathcal{L} =& \langle \ln p(\boldsymbol{y}|\boldsymbol{w})\rangle + \langle \ln p(\boldsymbol{w}|\boldsymbol{\alpha})\rangle + \langle \ln(\boldsymbol{\alpha})\rangle \\ & - \langle \ln q(\boldsymbol{w}) - \langle \ln q(\boldsymbol{\alpha})\rangle.\end{aligned} \tag{9}$$

The interested reader is redirected to [5] for the derivation and more details on the computation of Equation (9).

Motivated by extensive simulations we appose the following conjecture which suggests a simple and effective remedy to this problem.

Conjecture 1. *If $w_i^2 > z_i$ and the solution of Equation* (7) *results in three distinct real roots $0 < \tilde{\alpha}_i^1 < \tilde{\alpha}_i^2 < \tilde{\alpha}_i^3 < +\infty$ then $\tilde{\alpha}_i^1$ causes the variational lower bound $\mathcal{L}$ to increase the most.*

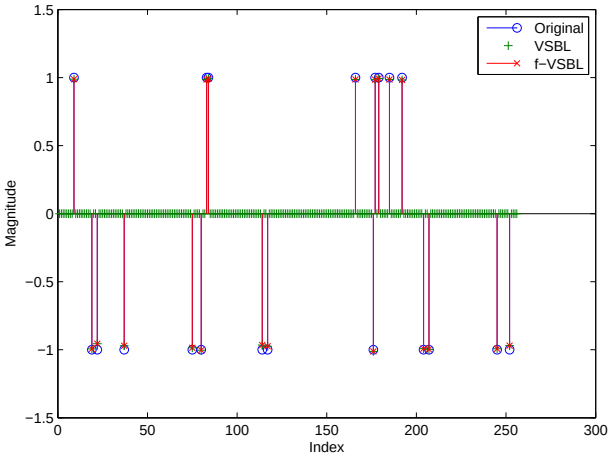
It is argued that in the case where three distinct real fixed points exist, the best choice as far as the lower bound is concerned is to select the smallest one in value. Conjecture 1 provides excellent empirical results and its proof is part of the authors' ongoing work.

3.2. Controlling complexity for superfluous parameters

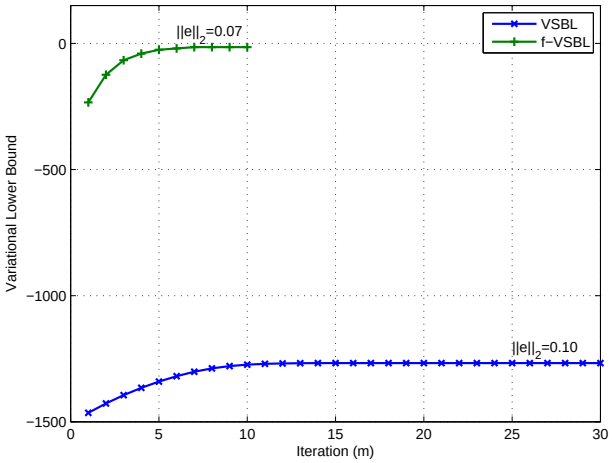
Improvements in terms of complexity can be achieved by reducing the number of parameters that need to be updated. Proposition 1 (Case 1) suggests that the parameter updates corresponding to component $x_i = 0$ cannot be pruned analytically since a valid fixed point is possible. Assessing its stability is analytically intractable and even then there is no guarantee that the fixed point $\tilde{\alpha}_i^*$ related to $x_i = 0$ will become unstable so as to disregard the corresponding parameter updates. A practical way is to update only the parameters for which $\tilde{\alpha}_i^*$ is above a certain threshold. In [6] it was proven that $w_i^2/z_i = SNR_i$, the signal-to-noise ratio for x_i. Our approach is to update only the parameters for which $\tilde{\alpha}_i^* > \sigma^2$.

4. EMPIRICAL RESULTS

The performance of the proposed algorithm namely the f-VSBL is assessed. At first we compare the performance of f-VSBL against VSBL in terms of sparse signal recovery, convergence speed (iteration count m) and reconstruction error $\boldsymbol{e}$. The entries of $\boldsymbol{\Phi} \in \mathbb{R}^{128\times 256}$ are drawn from $\mathcal{N}(0, 1/M)$. Signal $\boldsymbol{x}$ is a *zero-one* sparse signal with support set $\mathcal{T}$ chosen uniformly at random from $[1, N]$ with $|\mathcal{T}| = 20$ and



(a) f-VSBL returns an *exactly sparse* signal in contrast to VSBL.



(b) f-VSBL converges faster, at a higher lower bound with smaller error.

Fig. 1. Reconstruction performance for $\mathbf{\Phi} \in \mathbb{R}^{128\times256}$, $|\mathcal{T}| = 20$ and $\sigma^2 = 0.01$ for a *zero-one* sparse signal. The Gamma distribution parameters are set to $a_i = b_i = 0.1^3$.

$\sigma^2 = 0.01$. A single run of the algorithms was performed with $\boldsymbol{y}$, σ^2 as input. The sparsity level needs not be known, which is a powerful aspect of Bayesian methods. The Gamma parameters were set to $a_i = b_i = 0.1^3$ for all $i \in [1, N]$. In Figure 1(a) the original signal is shown versus the recovered. The f-VSBL does not suffer from the small amplitude components giving an exactly sparse signal. This causes a decrease in convergence speed as shown in 1(b) where f-VSBL converges in only 10 iterations. Convergence was assumed when the difference in the variational lower bound went below 0.1^8. The number of iterations was limited to 30 since VSBL in our tests took more than 700 iterations to converge. It is also shown that the f-VSBL achieves a significantly higher variational lower bound and higher reconstruction accuracy.

Table 1 compares the convergence speed and runtime (in seconds) of f-VSBL. In this scenario increasing problem sizes are considered, i.e., the design matrix $\mathbf{\Phi}$ is re-sampled at different sizes. The sparsity level and prior distribution strength are kept unchanged. It is evident that the proposed algorithm succeeds in recovering sparse signals under the *informative* assumption showcasing significantly reduced computational complexity and runtime.

For Table 2 we assume a stringent scenario. We consider a subset $\mathcal{S} \subset \mathcal{T}$ for which a stronger prior is employed expressing prior preference. We set $a_i = b_i = 0.1^5$ for $i \in \mathcal{T} - \mathcal{S}$ while the prior for $i \in \mathcal{S}$ varies as shown in Table 2. The algorithm is tested for different sizes of $\mathcal{S}$ against reconstruction error, recovered support set cardinality $\|\mathcal{T}'\|$ and iteration count. It is considered that $\|\mathcal{T}\| = 50$ while $\mathbf{\Phi} \in \mathbb{R}^{128\times256}$. Recovery using *uninformative* priors underperforms with $\|\boldsymbol{e}\|_2 = 0.55$ and $\mathcal{T}' = 71$. Table 2 shows that for adequately large $\mathcal{S}$ exact recovery is possible. By increasing the strength of the prior is it also possible to improve convergence speed.

Problem Size	Iterations (m)		Runtime (sec)	
	f-VSBL	VSBL	f-VSBL	VSBL
128×256	10	40	0.44	0.75
256×512	9	39	1.20	3.91
512×1024	8	38	6.63	25.28
1024×2048	9	38	62.2	142.88

Table 1. Comparison for $a_i = b_i = 0.1^3$, $|\mathcal{T}| = 20$, $\sigma^2 = 0.01$ and for increasing problem size.

Prior	$\|\mathcal{S}\| = 15$			$\|\mathcal{S}\| = 30$		
$a_i = b_i$	$\|\boldsymbol{e}\|_2$	$\|\mathcal{T}'\|$	m	$\|\boldsymbol{e}\|_2$	$\|\mathcal{T}'\|$	m
0.1^2	0.54	71	35	0.13	50	11
1	0.57	71	48	0.13	50	10
10^2	0.55	69	24	0.12	50	9
10^5	0.62	69	27	0.12	50	9

Table 2. Comparison for $\mathbf{\Phi} \in \mathbb{R}^{128\times256}$, $|\mathcal{T}| = 50$, $\sigma^2 = 0.01$ at different sizes of $\mathcal{S}$ against different prior strength.

5. CONCLUSION AND FUTURE WORK

Modifications have been proposed to VSBL that reduce complexity for when *informative priors* are considered. Complicated analytical expressions are avoided in the fixed point analysis. The empirical results support the theoretical claims. Ongoing work involves proving Conjecture 1 and extending the authors' previous work in [8, 9].

6. REFERENCES

[1] M.E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *The Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[2] A.C. Faul and M.E. Tipping, "Analysis of sparse Bayesian learning," in *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, pp. 383–390, 2002.

[3] M.E. Tipping and A.C. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Jan, 2003, vol. 1.

[4] S. Ji, Y. Xue and L. Carin, "Bayesian compressive sensing," *IEEE Transactions on Signal Processing*, vol. 56, no. 6, pp. 2346–2356, 2008.

[5] C.M. Bishop and M.E. Tipping, "Variational relevance vector machines," in *Proceedings of the Sixteenth conference on Uncertainty in Artificial intelligence*. Morgan Kaufmann Publishers Inc., pp. 46–53, 2000.

[6] D. Shutin, T. Buchgraber, S.R. Kulkarni, and H.V. Poor, "Fast variational sparse Bayesian learning with automatic relevance determination for superimposed signals," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6257–6261, 2011.

[7] T. Buchgraber, "Variational sparse Bayesian learning: Centralized and distributed processing," 2013, *PhD Thesis*.

[8] E. Karseras, K.K. Leung and W. Dai, "Tracking dynamic sparse signals using hierarchical Bayesian Kalman filters," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 6546–6550, 2013.

[9] E. Karseras, K.K. Leung and W. Dai, "Tracking dynamic sparse signals with Kalman filters: Framework and improved inference," in *Proceedings of the 10th International Conference on Sampling Theory and Applications (SampTA)*, pp.224–227, 2013.