Bayesian Compressed Sensing with Unknown Measurement Noise Level

Hansen, Thomas Lundgaard; Jørgensen, Peter Bjørn; Pedersen, Niels Lovmand; Manchón, Carles Navarro; Fleury, Bernard Henri

Published in:
Proc. 47th Asilomar Conference on Signals, Systems and Computers

DOI (link to publication from Publisher):
10.1109/ACSSC.2013.6810248

Publication date:
2013

Document Version
Early version, also known as pre-print

Bayesian Compressed Sensing with Unknown Measurement Noise Level

Thomas L. Hansen, Peter B. Jørgensen, Niels L. Pedersen, Carles Navarro Manchón and Bernard H. Fleury
Dept. of Electronic Systems, Aalborg University, Fr. Bajers Vej 7, DK-9220, Aalborg, Denmark
Email: {cnm,bfl}@es.aau.dk, Phone: +45 9940 8652

***Abstract*—In sparse Bayesian learning (SBL) approximate Bayesian inference is applied to find sparse estimates from observations corrupted by additive noise. Current literature only vaguely considers the case where the noise level is unknown a priori. We show that for most state-of-the-art reconstruction algorithms based on the fast inference scheme noise precision estimation results in increased computational complexity and reconstruction error. We propose a three-layer hierarchical prior model which allows for the derivation of a fast inference algorithm that estimates the noise precision with no complexity increase. Numerical results show that it matches or surpasses other algorithms in terms of reconstruction error.**

I. Introduction

Sparse signal representation from overcomplete dictionaries has found increasingly many applications in recent years, e.g. within compressed sensing [1], [2], machine learning [3] and channel estimation [4]. The canonical problem of interest can be formulated as

$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \mathbf{n}. \tag{1}$$

The $N \times 1$ observation vector $\mathbf{y}$ is corrupted by additive white Gaussian noise $\mathbf{n}$ with variance λ^{-1}. We seek a sparse representation $\mathbf{w}$ in the $N \times M$ dictionary $\mathbf{\Phi}$. The ith column $\boldsymbol{\phi}_i$ in the dictionary is the basis vector pertaining to the ith weight w_i. The number of observations N is much smaller than the number of basis vectors M, i.e. $N \ll M$. We consider both the case where $\mathbf{y}, \mathbf{\Phi}, \mathbf{w}$ and $\mathbf{n}$ are all real-valued and the case where they are all complex-valued.

The reconstruction algorithms already proposed can generally be classified into three categories: *a)* methods based on convex optimization, (e.g. [5], [6]), *b)* iterative constructive greedy algorithms (e.g. [7], [8]) and *c)* approaches based on Bayesian inference in sparsity-inducing probabilistic models. The latter are known as sparse Bayesian learning (SBL) approaches and they are the focus of this work.

Based on (1) the probabilistic model used in SBL for the observations $\mathbf{y}$ consists of a Gaussian likelihood function with mean $\mathbf{\Phi}\mathbf{w}$ and covariance matrix $\lambda^{-1}\mathbf{I}$

$$p(\mathbf{y}|\mathbf{w}) = \mathrm{N}\big(\mathbf{y}|\mathbf{\Phi}\mathbf{w}, \lambda^{-1}\mathbf{I}\big) \tag{2}$$

where $\mathbf{I}$ denotes the identity matrix. A prior probability density function (pdf) is specified for the noise precision λ. For the weight vector $\mathbf{w}$ a (possibly hierarchical) sparsity-inducing prior is selected. Through Bayesian inference a sparse estimate of the weights in $\mathbf{w}$ is obtained. The inference is typically done with an iterative scheme, because a closed form solution is infeasible.

Following a Bayesian approach the noise precision λ could be integrated out of the model (marginalized) prior to applying the inference scheme. As this is intractable in most cases, a point estimate of λ is obtained instead. The point estimate is either fixed at an initial rough estimate or updated periodically within the iterative inference. As seen in Section IV, SBL algorithms depend strongly on the accuracy of the point estimate of λ. Despite this fact, the estimation of λ has received surprisingly little attention in current literature.

A widely used SBL algorithm is the relevance vector machine (RVM) [3]. The original formulation of the RVM uses the expectation maximization (EM) algorithm [9] for inference. Inclusion of the estimation of noise precision in this iterative procedure is straightforward. The EM-based algorithm, however, requires a large number of iterations before convergence and has high computational cost per iteration. To improve on these aspects the 'fast inference scheme' for the RVM is introduced in [10]. This inference method, unlike EM, does not provide an integrated, simple way to estimate the noise precision. In fact, the computational cost of an iteration of the algorithm increases dramatically if the estimate of the noise precision is updated during that iteration. To circumvent this, it is proposed in [10] to only re-estimate λ once every few iterations while keeping its value fixed for the remaining iterations.

In [11] the fast inference scheme is used in combination with a hierarchical Laplace prior on $\mathbf{w}$. The resulting algorithm is shown to perform better than the RVM in terms of mean-squared error (MSE) of the weights. In the numerical results the noise precision is kept fixed through all iterations at an initial estimate $\hat{\lambda} = 0.01\,||\mathbf{y}||_2^2$. It is argued that the noise precision cannot be estimated in practice, as the fast inference scheme produces unreliable estimates in the first few iterations.

In [12] a hierarchical model of the Bessel K prior is presented. Algorithms resulting from applying the fast inference scheme to the Bessel K prior model are shown to perform extremely well, but they also suffer from higher computational complexity when estimating the noise precision.

In [13] a slightly modified version of the model used in the RVM is presented. In this model it is tractable to integrate out the noise precision and an estimate of λ is thus not required for inference. Our numerical investigations indicates that this algorithm has performance similar to that of the RVM.

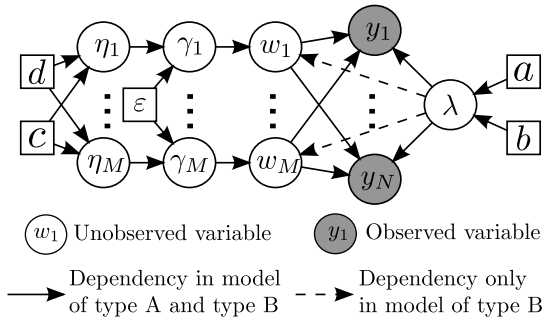

Fig. 1: Bayesian network of probabilistic model A and B.

Density	Model A	Model B
Observations, $p(\mathbf{y}\|\mathbf{w},\lambda)$	$\mathrm{N}\big(\mathbf{y}\|\mathbf{\Phi}\mathbf{w},\lambda^{-1}\mathbf{I}\big)$	
Prior on λ, $p(\lambda)$	$\mathrm{Ga}(\lambda\|a,b)$	
Layer 1 on weights, $p(\mathbf{w}\|\boldsymbol{\gamma})$	$\mathrm{N}(\mathbf{w}\|\mathbf{0},\mathbf{\Gamma})$	$\mathrm{N}\big(\mathbf{w}\|\mathbf{0},\lambda^{-1}\mathbf{\Gamma}\big)$
Layer 2 on weights, $p(\boldsymbol{\gamma}\|\boldsymbol{\eta})$	$\prod_{i=1}^{M}\mathrm{Ga}(\gamma_i\|\varepsilon,\eta_i)$	
Layer 3 on weights, $p(\boldsymbol{\eta})$	$\prod_{i=1}^{M}\mathrm{Ga}(\eta_i\|c,d)$	

Definitions: We define the vector $\boldsymbol{\eta}=[\eta_1,\ldots,\eta_M]^{\mathrm{T}}$ and the diagonal matrix $\mathbf{\Gamma}=\mathrm{diag}(\boldsymbol{\gamma})$ with the vector $\boldsymbol{\gamma}=[\gamma_1,\ldots,\gamma_M]^{\mathrm{T}}$.
The multivariate normal density is parameterized to encompass both the real ($\rho=\frac{1}{2}$) and complex ($\rho=1$) case:

$$\mathrm{N}(\mathbf{x}|\boldsymbol{\mu},\mathbf{\Sigma})=\Big(\frac{\rho}{\pi}\Big)^{\rho\dim(\mathbf{x})}|\mathbf{\Sigma}|^{-\rho}\exp\Big(-\rho(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{H}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\Big)$$

where $(\cdot)^{\mathrm{H}}$ denotes the (Hermitian) matrix transpose.
The gamma pdf with shape $\alpha>0$ and rate $\beta>0$ is

$$\mathrm{Ga}(x|\alpha,\beta)=\beta^{\alpha}\Gamma(\alpha)^{-1}x^{\alpha-1}\exp\left(-\beta x\right)$$

where $\Gamma(\alpha)$ is the gamma function.

TABLE I: Probability densities in probabilistic model A and B.

From the above discussion, it is clear estimation of the noise precision is not straightforward, that many different approaches have been proposed and further investigation is needed to identify the most promising options.

In this paper we present an algorithm that includes estimation of the noise precision in the inference framework without any increase in the computational complexity. We propose a generalization of the hierarchical prior model in [13] from which a novel fast inference algorithm is derived. A comparison is made with the hierarchical model in [12]. Unlike many other SBL algorithms, the performance of the proposed algorithm is the same whether the noise is estimated or fixed to its true value.

The paper is organized as follows; in Section II we present the two investigated probabilistic models and relate them to models currently used in the literature. In Section III we derive a novel sparse estimation algorithm by applying the fast inference scheme to our proposed model. Results of our numerical investigation are presented and discussed in Section IV and conclusions follow in Section V.

II. Probabilistic Modelling

In this paper we investigate two different probabilistic models denoted as model A and model B, respectively. Fig. 1 shows the Bayesian network of the two models. Table I shows the pdfs used. Model A is presented in [12]. We propose model B as a generalization of the models in [13] and [14]. The sole difference between model A and B lies in the specification of the variance in the first layer on the weights. For model A the variance of w_i is specified by γ_i, while for model B it is given by $\gamma_i\lambda^{-1}$. In model B each γ_i can be interpreted as a signal-to-noise ratio (SNR) for the basis vector $\boldsymbol{\phi}_i$.

Notice how the weights $\mathbf{w}$ are modelled through a three-layer (3L) hierarchical prior specification. We also refer to two-layer (2L) versions of the models, where the prior on $\boldsymbol{\eta}$ is disregarded and η_i, $i=1,\ldots,M$ is instead considered as parameters of the model.

The model used to derive the RVM [3], [10] is different from the models presented above. It can however be derived from model A by selecting a flat (improper) prior on $\gamma_i, i=1,\ldots,M$ (as done in [15]). Similarly, the model used in [13] is obtained by imposing a flat prior on γ_i in the 2L version of model B. Therefore, the algorithm in [13] can be considered an analogue to the RVM. The flat prior on γ_i is in both models obtained by selecting $\varepsilon=1$ and letting $\eta_i\to 0$. The Laplace prior model presented in [11] is obtained with an exponential prior on γ_i in model A, which is realized when $\varepsilon=1$. Notice that in [11] $\eta_i=\eta,\ \forall\ i$.

III. Bayesian Inference

Based on the presented probabilistic model we derive an estimator of the weights. In the following we apply the fast inference scheme to the 3L version of model B and refer to [12] for corresponding algorithms based on the 2L and 3L versions of model A. We follow the conventional approach within SBL (e.g. [3], [10]–[12], [15]) and obtain estimates ($\hat{\boldsymbol{\gamma}}$, $\hat{\lambda}$, $\hat{\boldsymbol{\eta}}$) of the hyperparameters ($\boldsymbol{\gamma}$, λ, $\boldsymbol{\eta}$). The estimate of $\mathbf{w}$ is then obtained as the mode of $p(\mathbf{w}|\mathbf{y},\hat{\boldsymbol{\gamma}},\hat{\lambda})$.

Note that $p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\lambda)\propto p(\mathbf{y}|\mathbf{w},\lambda)p(\mathbf{w}|\boldsymbol{\gamma},\lambda)$ is the Gaussian pdf given by

$$p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\lambda)=\mathrm{N}\big(\mathbf{w}|\boldsymbol{\mu},\lambda^{-1}\mathbf{\Sigma}\big) \tag{3}$$

where

$$\boldsymbol{\mu}=\mathbf{\Sigma}\mathbf{\Phi}^{\mathrm{H}}\mathbf{y},\qquad \mathbf{\Sigma}=\left(\mathbf{\Phi}^{\mathrm{H}}\mathbf{\Phi}+\mathbf{\Gamma}^{-1}\right)^{-1}. \tag{4}$$

Hence, the mode of $p(\mathbf{w}|\mathbf{y},\boldsymbol{\gamma},\lambda)$ coincides with $\boldsymbol{\mu}$ in (4).

The fast inference scheme estimates the hyperparameters $(\boldsymbol{\gamma},\lambda,\boldsymbol{\eta})$ based on iterative maximization of the posterior pdf

$$p(\boldsymbol{\gamma},\lambda,\boldsymbol{\eta}|\mathbf{y})\propto p(\mathbf{y}|\boldsymbol{\gamma},\lambda)p(\boldsymbol{\gamma}|\boldsymbol{\eta})p(\boldsymbol{\eta})p(\lambda), \tag{5}$$

where

$$p(\mathbf{y}|\boldsymbol{\gamma},\lambda)=\int p(\mathbf{y}|\mathbf{w},\lambda)p(\mathbf{w}|\boldsymbol{\gamma},\lambda)\,\mathrm{d}\mathbf{w}=\mathrm{N}\big(\mathbf{y}|\mathbf{0},\lambda^{-1}\mathbf{B}\big) \tag{6}$$

with

$$\mathbf{B}=\mathbf{I}+\mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^{\mathrm{H}}. \tag{7}$$

The matrix $\mathbf{B}$ can be decomposed as

$$\mathbf{B}=\mathbf{I}+\sum_{k\neq i}\boldsymbol{\phi}_k\gamma_k\boldsymbol{\phi}_k^{\mathrm{H}}+\boldsymbol{\phi}_i\gamma_i\boldsymbol{\phi}_i^{\mathrm{H}}=\mathbf{B}_{-i}+\boldsymbol{\phi}_i\gamma_i\boldsymbol{\phi}_i^{\mathrm{H}}. \tag{8}$$

From Woodbury's matrix inversion identity and the matrix determinant lemma, we get

$$\mathbf{B}^{-1} = \mathbf{B}_{-i}^{-1} - \frac{\mathbf{B}_{-i}^{-1}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^{\mathrm{H}}\mathbf{B}_{-i}^{-1}}{\gamma_i^{-1} + \boldsymbol{\phi}_i^{\mathrm{H}}\mathbf{B}_{-i}^{-1}\boldsymbol{\phi}_i}, \tag{9}$$

$$|\mathbf{B}| = |\mathbf{B}_{-i}|\left(1 + \gamma_i\boldsymbol{\phi}_i^{\mathrm{H}}\mathbf{B}_{-i}^{-1}\boldsymbol{\phi}_i\right). \tag{10}$$

Taking the log of the posterior in (5) and inserting (9) and (10) yields

$$\begin{aligned}\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\eta},\lambda) &= \log p(\boldsymbol{\gamma},\lambda,\boldsymbol{\eta}|\mathbf{y}) \\ &= (\rho N + a - 1)\log\lambda - \rho\log|\mathbf{B}_{-i}| - \rho\log(1+\gamma_i s_i) \\ &+ \sum_{j=1}^{M}\left((\varepsilon-1)\log\gamma_j + (\varepsilon+c-1)\log\eta_j - (\gamma_j+d)\eta_j\right) \\ &+ \rho\lambda\left(\frac{|q_i|^2}{\gamma_i^{-1}+s_i} - g_i\right) + \text{const.}\end{aligned} \tag{11}$$

where we have defined the quantities

$$s_i = \boldsymbol{\phi}_i^{\mathrm{H}}\mathbf{B}_{-i}^{-1}\boldsymbol{\phi}_i, \quad q_i = \boldsymbol{\phi}_i^{\mathrm{H}}\mathbf{B}_{-i}^{-1}\mathbf{y}, \quad g_i = \mathbf{y}^{\mathrm{H}}\mathbf{B}_{-i}^{-1}\mathbf{y} + \frac{b}{\rho}. \tag{12}$$

The decomposition (8) enables maximization of (11) with respect to one set of hyperparameters (γ_i, η_i). We could now proceed by maximizing sequentially with respect to γ_i and then η_i. However, numerical results show that maximizing jointly with respect to (γ_i, η_i) reduces the number of required iterations to reach convergence by more than a factor of two in most scenarios. We choose $\varepsilon = 1$, such that the second layer is governed by an exponential density. This simplifies the derivations and yields algorithms with good performance as our numerical results show. Through differentiation and substitution, the stationary points of (11) with respect to (γ_i, η_i) are found by solving

$$\begin{aligned}\gamma_i^2 s_i^2(\rho+c) + \gamma_i(2s_i c + d\rho s_i^2 + \rho(s_i - \lambda|q_i|^2)) \\ + c + d\rho(s_i - \lambda|q_i|^2) = 0.\end{aligned} \tag{13}$$

By analyzing (13) we realize that; *a)* when no positive root of (13) exists, the global maximizer of (11) on $\mathbb{R}^+$ is at $\gamma_i = 0$, *b)* in the case of one positive root, this root is a global maximizer on $\mathbb{R}^+$ and *c)* when there are two positive roots, the largest is a local (in some cases global) maximizer. However, empirical results show that discarding solutions obtained from case *c)* increases the reconstruction performance of the algorithms. The solutions obtained from this case have been observed to give very small values of γ_i compared to those obtained in case *b)*. As this results in small values for the corresponding w_i it intuitively makes sense to force those γ_i to zero. Only using the maximizers from case *a)* and *b)*, the update expression reads

$$\begin{aligned}\hat{\gamma}_i &= \begin{cases}\frac{\rho(\hat{\lambda}|q_i|^2 - s_i) - 2s_i c - d\rho s_i^2 + \sqrt{\Delta_i}}{2s_i^2(\rho+c)} & \text{if } \hat{\lambda}|q_i|^2 - s_i > \frac{c}{d\rho} \\ 0 & \text{otherwise}\end{cases} \\ \hat{\eta}_i &= \frac{c}{\hat{\gamma}_i + d}\end{aligned} \tag{14}$$

	Model A	Model B
Fixed $\hat{\lambda}$	$\mathcal{O}(MN)$	$\mathcal{O}(MN)$
Updating $\hat{\lambda}$	$\mathcal{O}(MN\hat{S})$	$\mathcal{O}(MN)$

TABLE II: Computational cost of each iteration using the fast inference scheme. It is assumed that $\hat{S} \leq N \leq M$, where $\hat{S}$ is the number of nonzero components in $\hat{\mathbf{w}}$ in the given iteration.

where $\Delta_i = (2s_i c + d\rho s_i^2 + \rho(s_i - \hat{\lambda}|q_i|^2))^2 - 4s_i^2(\rho+c)(c + d\rho(s_i - \hat{\lambda}|q_i|^2))$. Maximization of (11) with respect to λ leads to the following update expression for the noise precision estimate

$$\hat{\lambda} = \frac{N + \frac{a-1}{\rho}}{\mathbf{y}^{\mathrm{H}}\mathbf{B}^{-1}\mathbf{y} + \frac{b}{\rho}}. \tag{15}$$

The fast inference scheme starts with an 'empty' model by setting all values in $\hat{\boldsymbol{\gamma}}$ to zero. The algorithm proceeds by iteratively selecting a basis vector for which $(\hat{\gamma}_i, \hat{\eta}_i)$ are recalculated according to (14). Depending on the selected index i, an update can consist of addition or deletion of a basis vector or re-estimation of the parameters corresponding to a basis vector already included in the model. In our implementation we, similarly to [10]–[13], choose to update the pair $(\hat{\gamma}_i, \hat{\eta}_i)$ which results in the largest increase in $\mathcal{L}(\boldsymbol{\gamma},\boldsymbol{\eta},\lambda)$ at each particular iteration. In each iteration the noise precision is re-estimated through (15).

Algorithms derived from model B can use the update formulas in [13] to update $\boldsymbol{\Sigma}, \boldsymbol{\mu}$ and $(s_i, q_i, g_i)\,\forall\, i$ in each iteration at reduced computational cost. Equivalent update formulas can be found in [10] for inference in model A. A key difference between the two models is that the update formulas for model A are only valid when the noise precision estimate $\hat{\lambda}$ is held fixed between two consecutive iterations, whereas they are applicable in all iterations in model B. When using model A, the quantities must be computed using their definitions when $\hat{\lambda}$ is updated. The computational complexity in each scenario is summarized in Table II. In the original work [10] it is proposed to only update the noise precision estimate after every fifth iteration.

In model B it also becomes tractable to marginalize out the noise precision as done for a special case in [13]. When marginalizing λ, the derivations follow the ones above and due to space limitations they are not presented here. The increased tractability in model B arises because the expressions (4) and (7) no longer depend on the noise precision λ. When deriving algorithms based on the fast inference scheme, this decoupling is exploited. We note that similar benefits arise when using other approaches for inference in model B, e.g. [16].

IV. Numerical Results

In this section we assess the performance of different algorithms through numerical simulations. Table III lists the SBL algorithms we consider. The A-RVM [10] and A-Laplace [11] are established algorithms within SBL and are considered as important references. B-RVM is the algorithm proposed in

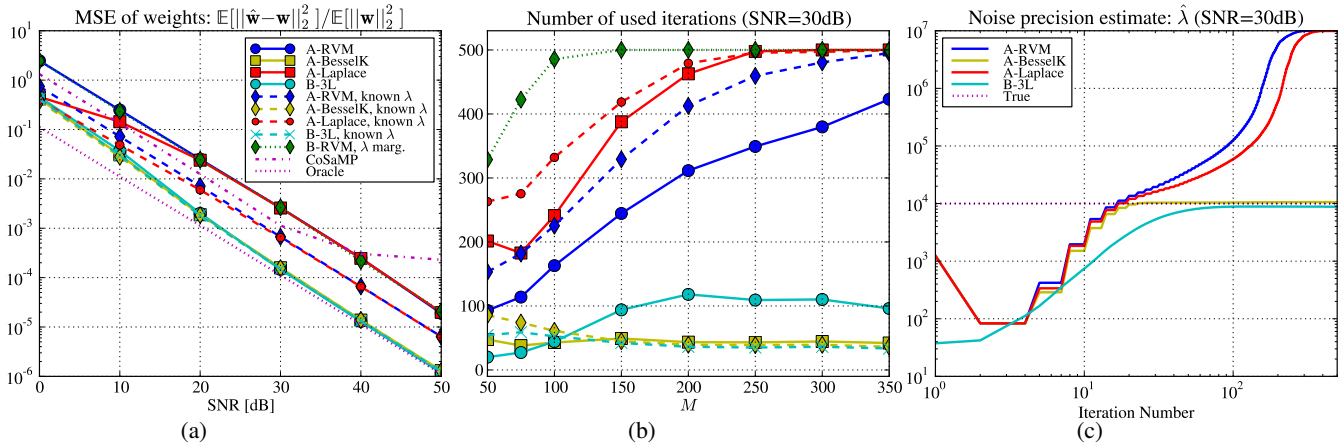

Fig. 2: Performance comparison of different sparse estimation algorithms with known and unknown noise precision. The Python based simulation code used to generate these plots, can be found at `http://www.es.aau.dk/navcom/sbl_index`. Note that the legend in (a) is also valid for (b).

Algorithm	Parameters	References
A-RVM	$\varepsilon = 1, \eta_i = 0\,\forall i$	[10], [15]
A-BesselK	$\varepsilon = 0, \eta_i = 1\,\forall i$	[12]
A-Laplace	$\varepsilon = 1, c = d = 0, \eta_i = \eta\,\forall i$	[11]
B-RVM, λ marg.	$\varepsilon = 1, \eta_i = 0\,\forall i$	[13]
B-3L	$\varepsilon = 1, c = 0.5, d = 0.1$	

TABLE III: List of the investigated SBL algorithms. All algorithms use $a = 1, b = 0$, i.e. a 'flat' prior is used for the noise precision λ.

[13]. It is a direct analogue to A-RVM, but uses model B and the noise precision is marginalized out. The A-BesselK algorithm is presented in [12]. The B-3L is the algorithm proposed in this paper. We omit results for the 2L version of model B (B-2L), as we have been unable to select a single parameter value for η_i which performs well for all SNR values. The value of the parameters $c = 0.5$ and $d = 0.1$ have been chosen empirically to optimize the reconstruction performance. Notice that $\frac{c}{d}$ affects the sparsity of the obtained estimates, with larger $\frac{c}{d}$ producing estimates with fewer non-zero components.

For easier comparison we use a 'flat' prior for the noise precision in all algorithms as in the RVM. As some of the algorithms severely overestimate the noise precision, we limit the noise precision estimate to 10^7 to avoid numerical instabilities. For the algorithms using model A, the noise precision estimate is only updated every third iteration. All algorithms terminate when $||\hat{\mathbf{w}}_n - \hat{\mathbf{w}}_{n-1}||_\infty < 10^{-8}$ with $\hat{\mathbf{w}}_n$ and $\hat{\mathbf{w}}_{n-1}$ denoting the estimate of $\mathbf{w}$ in the current and previous iteration, respectively. In addition we limit the maximum number of iterations to 500.

We include the oracle estimator as a reference. This estimator knows the support of $\mathbf{w}$ and computes a least-squares estimate of the nonzero entries in $\mathbf{w}$. The CoSaMP [8] algorithm is a state-of-the-art non-Bayesian reconstruction algorithm from compressed sensing and is also included as a reference.

We use a generic simulation scenario and obtain the observations in accordance with (1). Each simulation uses a randomly generated dictionary $\mathbf{\Phi}$ with entries independently and identically distributed according to a zero-mean normal distribution with variance $1/N$. The number of nonzero weights is binomially distributed with mean 15. The location of the nonzero weights is uniformly distributed and the value of each nonzero entry is sampled from a standard normal distribution. Unless otherwise stated, $M = 300$ and the number of observations is $N = \frac{M}{2}$. The SNR is given by

$$\text{SNR} = \frac{\mathbb{E}\left[||\mathbf{\Phi}\mathbf{w}||_2^2\right]}{\mathbb{E}\left[||\mathbf{n}||_2^2\right]} = \frac{\lambda \bar{S}}{N} \tag{16}$$

where $\bar{S}$ is the average number of nonzero entries in $\mathbf{w}$. In the considered scenario $\mathbf{y}$, $\mathbf{\Phi}$, $\mathbf{w}$ and $\mathbf{n}$ are all real-valued. The initial noise precision estimate is chosen as $\hat{\lambda} = \frac{100}{\text{var}(\mathbf{y})}$, where $\text{var}(\mathbf{y})$ denotes the sample variance of $\mathbf{y}$. All results are averaged over 100 Monte Carlo simulations.

The MSE of the weight vector estimate is shown versus the SNR in Fig. 2(a). Notice that for A-RVM and A-Laplace the MSE increases when the noise precision is estimated compared to when it is known. As depicted in Fig. 2(c) these algorithms keep increasing their noise precision estimate over iterations and never reach convergence. We have observed that the algorithms keep adding basis vectors to the model (not shown here) and obtain a non-sparse solution, i.e. they do overfitting. In [11] it is argued that this problem is caused by the construction of the fast inference scheme, as it starts with an empty model and therefore produces unreliable noise precision estimates in the first few iterations. However, our simulations show that other SBL algorithms (A-BesselK and B-3L) are able to cope with unknown observation noise level without any degradation in reconstruction performance.

In Fig. 2(b) the number of used iterations is plotted versus M (note that $N = \frac{M}{2}$ and N is therefore also varied). The algorithms that have a tendency to produce non-sparse estimates (A-RVM and A-Laplace with estimated and known λ and B-RVM) require more iterations when M increases, as there are more candidate basis vectors to be added. The number of used iterations for A-BesselK and the proposed B-3L does not increase with M or N for $M \geq 150$. When λ is known, A-BesselK and B-3L have the same computational complexity per iteration and use the same number of iterations. For unknown λ the computational complexity per iteration is higher for A-BesselK compared to B-3L ($\mathcal{O}(MN\hat{S})$ versus $\mathcal{O}(MN)$). B-3L is however seen to require a larger number of iterations before convergence. In summary, the proposed B-3L algorithm shows as good performance as the best state-of-the-art SBL estimators and is more computationally efficient per iteration when estimating the noise precision, at the cost of a higher number of iterations required before convergence.

V. Conclusion

In this paper we have investigated Bayesian compressed sensing methods and how they deal with unknown observation noise levels. We have shown that the reconstruction performance of state-of-the-art algorithms employing the fast inference scheme by Tipping and Faul [10] is degraded when the noise precision needs to be estimated. Both the fast RVM [10] and the algorithm proposed in [11] using a Laplace prior model overestimate the noise precision and produce non-sparse estimates. This is a shortcoming of the used prior model. Using the 2-layer prior model in [12], which favours more sparse solutions, yields an algorithm that produces an unbiased estimate of the noise precision and favorable reconstruction performance of the weights. Estimating the noise in the above mentioned algorithms, however, increases the computational complexity of the algorithms.

Through a modified probabilistic model inspired by the model in [13] it becomes possible to either estimate or marginalize out the noise precision, while preserving the low computational complexity of the fast inference scheme. On this basis we have proposed a novel sparse estimation algorithm using a three-layer probabilistic model. The reconstruction performance of this algorithm is on par with current state-of-the-art algorithms. Conversely to existing algorithms, our proposed algorithm retains the low computational complexity per iteration of the fast inference scheme when the noise precision is estimated.

References

[1] E. J. Candes and M. B. Wakin, "An introduction to compressive sampling," *IEEE Signal Processing Magazine*, vol. 25, no. 2, pp. 21–30, Mar. 2008.

[2] D. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[3] M. E. Tipping, "Sparse Bayesian learning and the Relevance Vector Machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, Jun. 2001.

[4] W. Bajwa, A. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[5] R. Tibshirani, "Regression shrinkage and selection via the LASSO," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[6] S. S. Chen, D. L. Donoho, and M. A. Sunders, "Atomic decomposition by Basis Pursuit," *SIAM Journal on Scientific Computing*, vol. 20, no. 1, pp. 33–61, 1999.

[7] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad, "Orthogonal matching pursuit: recursive function approximation with application to wavelet decomposition," in *The 27th Asilomar Conference on Signals, Systems and Computers*, vol. 1, Nov. 1993, pp. 40–44.

[8] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society. Series B*, vol. 39, no. 1, 1977.

[10] M. E. Tipping and A. Faul, "Fast marginal likelihood maximisation for sparse Bayesian models," in *Proceedings of the International Workshop on Artificial Intelligence and Statistics*, 2003, pp. 3–6.

[11] S. Babacan, R. Molina, and A. Katsaggelos, "Bayesian compressive sensing using Laplace priors," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 53–63, Jan. 2010.

[12] N. L. Pedersen, D. Shutin, C. N. Manchón, and B. H. Fleury, "Sparse estimation using bayesian hierarchical prior modeling for real and complex models," submitted to *IEEE Transactions on Signal Processing*, 2013, arXiv:1108.4324.

[13] S. Ji, D. Dunson, and L. Carin, "Multitask compressive sensing," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 92–106, Jan. 2009.

[14] T. Park and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, pp. 681–686, Jun. 2008.

[15] D. P. Wipf and B. D. Rao, "Sparse Bayesian learning for Basis Selection," *IEEE Transactions on Signal Processing*, vol. 52, no. 8, pp. 2153–2164, Aug. 2004.

[16] D. Shutin, T. Buchgraber, S. R. Kulkarni, and H. V. Poor, "Fast adaptive variational sparse bayesian learning with automatic relevance determination," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2180–2183.