BAYESIAN COMPRESSED SENSING: IMPROVING INFERENCE

Evripidis Karseras, Kin Leung, and Wei Dai

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

ABSTRACT

In this paper we present a set of theoretical results regarding inference algorithms for hierarchical Bayesian networks. More specifically we focus on a specific type of networks which result in highly sparse models for the input. Bayesian inference in these networks usually is based on optimising a non-convex cost function of the model parameters. We extend previous work done in this field by providing some global performance guarantees regarding this cost function. This is the starting point for redesigning the aforementioned algorithms by employing results from well known sparse reconstruction techniques. This contribution comes in the form of three theorems. The end result is a new view of the Bayesian sparse reconstruction problem.

Index Terms— Hierarchical, Bayesian, Subspace, Pursuit

1. INTRODUCTION

Sparse Bayesian Learning (SBL) was introduced in [1] as a hierarchical Bayesian network which results in highly sparse models for the given input. This is also known as the basic element of the Relevance Vector Machine (RVM). Even-though the model introduces a relatively large number of parameters, the resulting models do not suffer from over-learning but are surprisingly sparse. This is possible because of the nature of the prior distributions employed in this type of Bayesian networks. In the core of the inference procedure lies the optimisation of a non-convex cost function, with respect to the model parameters. The optimisation of this cost function is realised via an iterative algorithm and upon convergence most of the model parameters become irrelevant for describing the given dataset.

The starting point for this work is the employment of this technique for compressed sensing and basis selection as presented in [2] and [3] respectively. We provide a theoretical analysis of the cost function and generalise results from [3]. This analysis serves as the connecting link between already established sparse reconstruction algorithms and the Bayesian approach which provides a more versatile output, i.e full distributions rather than point estimates. Taking the progress one step further, we demonstrate how it is possible to improve the inference algorithms in [4] with attributes from well-known sparse reconstruction algorithms. Two algorithms are provided that exhibit this improved behaviour. The results are also verified empirically in Section 6.

2. SPARSE BAYESIAN LEARNING

Lets consider the problem of reconstructing sparse signal $\boldsymbol{x}$ from its noisy measurements:

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

where $\boldsymbol{y} \in \mathbb{R}^m$, $\boldsymbol{\Phi} \in \mathbb{R}^{m\times n}$, $\boldsymbol{x} \in \mathbb{R}^n$ and measurement noise is assumed to be white Gaussian with variance σ^2. In SBL a hierarchy of distributions is employed to model this setting. More specifically each component x_i is assumed to be dependent on a separate hyper-parameter α_i. Each α_i follows a suitably chosen distribution with it being *uninformative*, i.e uniform. This results in the prior distribution:

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

where $\boldsymbol{A} = \mathrm{diag}\left(\boldsymbol{\alpha}\right) = \mathrm{diag}\left(\left[\alpha_1, \cdots, \alpha_n\right]_t\right)$, and the hyper-parameters α_i are unknown and have to be learned from $\boldsymbol{y}$. By driving $\alpha_i = +\infty$ it means that $p\left(x_i|\alpha_i\right) = \mathcal{N}\left(0,0\right)$; hence it is certain that $x_i = 0$. What remains is to find the maximum likelihood solution of $\boldsymbol{\alpha}$ for the given observation vector $\boldsymbol{y}$. The explicit form of the likelihood function $p\left(\boldsymbol{y}|\boldsymbol{\alpha}, \sigma^2\right)$ was derived in [1]:

$$\mathcal{L}(\boldsymbol{\alpha}) = \log|\boldsymbol{C}| + \boldsymbol{y}^T\boldsymbol{C}^{-1}\boldsymbol{y} \tag{2}$$

where $\boldsymbol{C} = \sigma^2\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{A}^{-1}\boldsymbol{\Phi}^T$. A set of fast algorithms to estimate $\boldsymbol{\alpha}$ were proposed in [4].

The closed form formula for the optimal value of a single hyper-parameter is:

$$\alpha_i = \frac{s_i^2}{q_i^2 - s_i} \tag{3}$$

The authors would like to acknowledge the European Commission for funding SmartEN ITN (Grant No. 238726) under the Marie Curie ITN FP7 programme. This work is partially supported by the University Defence Research Centre (UDRC), UK.

978-1-4799-1043-4/13/$31.00 ©2013 IEEE

where $s_i = \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1} \boldsymbol{\phi}_i$ and $q_i = \boldsymbol{\phi}_i^T \boldsymbol{C}_{-i}^{-1} \boldsymbol{y}$. Subscript $-i$ denotes subtracting the contribution from vector $\boldsymbol{\phi}_i$.

3. REVISED COST FUNCTION

It is easy to verify that by driving $\sigma^2 = 0$ matrix $\boldsymbol{C}$ becomes badly conditioned and thus the optimisation algorithm for cost function (2) performs poorly. To remedy this problem we provide Theorem 1.

Theorem 1. *For any given $\boldsymbol{\alpha}$, define the set $\mathcal{I} \triangleq \{1 \leq i \leq n : \; 0 < \alpha_i < \infty\}$. Then it holds that:*

$$\lim_{\sigma^2 \to 0} \sigma^2 \mathcal{L}(\boldsymbol{\alpha}) = \left\| \boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y} \right\|_2^2,$$

where $\boldsymbol{\Phi}_\mathcal{I}$ is a sub-matrix of $\boldsymbol{\Phi}$ formed by the columns indexed by $\mathcal{I}$, and $\boldsymbol{\Phi}_\mathcal{I}^\dagger$ denotes the pseudo-inverse of $\boldsymbol{\Phi}_\mathcal{I}$.

Furthermore, if $|\mathcal{I}| < m$ and $\boldsymbol{y} \in \operatorname{span}(\boldsymbol{\Phi}_\mathcal{I})$, then $\mathcal{L}(\boldsymbol{\alpha}) \to -\infty$ and $\sigma^2 \mathcal{L}(\boldsymbol{\alpha}) \to 0$ as $\sigma^2 \to 0$.

Proof: Consider the cost function as given by Equation 2. In order to derive the properly scaled version of the cost function the determinant and the inverse of matrix $\boldsymbol{C}$ are rewritten likewise:

$$\log |\boldsymbol{C}| = -m \log \left|\sigma^{-2} \boldsymbol{I}\right| + \log \left| \boldsymbol{I} + \sigma^{-2} \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{A}_\mathcal{I}^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \right|$$
$$\boldsymbol{C}^{-1} = \sigma^{-2} \boldsymbol{I} - \sigma^{-2} \boldsymbol{\Phi}_\mathcal{I} \left(\sigma^2 \boldsymbol{A}_\mathcal{I} + \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I} \right)^{-1} \boldsymbol{\Phi}_\mathcal{I}^T$$

where in the first equation a common property for the determinant was used and in the second equation the Woodbury matrix inversion formula was employed. Now the cost function becomes:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\alpha}) &= -m \log \sigma^{-2} + \log \left| \boldsymbol{I} + \sigma^{-2} \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{A}_\mathcal{I}^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \right| \\ &\quad + \sigma^{-2} \boldsymbol{y}^T \left(\boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \left(\sigma^2 \boldsymbol{A}_\mathcal{I} + \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I} \right)^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{y} \right) \\ &= o(\sigma^{-2}) + \sigma^{-2} \boldsymbol{y}^T \left(\boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \left(\sigma^2 \boldsymbol{A}_\mathcal{I} + \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I} \right)^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{y} \right) \end{aligned}$$

hence in the case noise variance σ^2 approaches zero:

$$\begin{aligned} \lim_{\sigma^2 \to 0} \sigma^2 \mathcal{L}(\boldsymbol{\alpha}) &= \boldsymbol{y}^T \left(\boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \left(\boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I} \right)^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{y} \right) \\ &= \boldsymbol{y}^T (\boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y}) \end{aligned}$$

where in the last step it is assumed that $|\mathcal{I}| < m$ and the expression for the pseudo-inverse is used. Now let $\boldsymbol{y}_p = \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y}$ and $\boldsymbol{y}_r = \boldsymbol{y} - \boldsymbol{y}_p$ denote the projection of $\boldsymbol{y}$ on the span of $\boldsymbol{\Phi}_\mathcal{I}$ and the residual signal respectively. The following holds:

$$\langle \boldsymbol{y}, \boldsymbol{y}_r \rangle = \langle \boldsymbol{y}_p + \boldsymbol{y}_r, \boldsymbol{y}_r \rangle = \| \boldsymbol{y}_r \|_2^2$$

since:

$$\begin{aligned} \langle \boldsymbol{y}_p, \boldsymbol{y}_r \rangle &= (\boldsymbol{\Phi}_\mathcal{I} \boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y})^T (\boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I} \boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y}) \\ &= (\boldsymbol{\Phi}_\mathcal{I}^\dagger \boldsymbol{y}) \left(\boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{y} - \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I} (\boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{\Phi}_\mathcal{I})^{-1} \boldsymbol{\Phi}_\mathcal{I}^T \boldsymbol{y} \right) = 0 \end{aligned}$$

Finally the scaled cost function can be written as:

$$\lim_{\sigma^2 \to 0} \mathcal{L}(\boldsymbol{\alpha}) = \| \boldsymbol{y}_r \|_2^2$$

which proves the first part of the theorem.

For the second part, assume that $\boldsymbol{y} \in \operatorname{span}(\Phi_\mathcal{I})$ and that $|\mathcal{I}| < m$. Then in the case of zero noise variance σ^2:

$$|\boldsymbol{C}| = |\boldsymbol{\Phi}_\mathcal{I} \boldsymbol{A}_\mathcal{I}^{-1} \boldsymbol{\Phi}_\mathcal{I}^T| = 0$$

since $\mathrm{rank}(\Phi_\mathcal{I} \mathrm{A}_\mathcal{I}^{-1} \Phi_\mathcal{I}^\mathrm{T}) < \mathrm{m}$ and $\boldsymbol{C} \in \mathbb{R}^{m \times m}$. Hence:

$$\lim_{\sigma^2 \to 0} \mathcal{L}(\boldsymbol{\alpha}) = -\infty$$

On the contrary, for the scaled cost function and in the limit of zero noise, the following holds:

$$\lim_{\sigma^2 \to 0} \sigma^2 \mathcal{L}(\boldsymbol{\alpha}) = 0$$

which completes the proof of Theorem 1.

SBL has previously been analysed in [3] for basis selection. More specifically it had been proven that a maximally sparse solution of $\boldsymbol{y} = \boldsymbol{\Phi x}$ is attained at the global minimum of the cost function. However, the analysis did not specify the conditions to avoid local minima. In Theorem 1 we provide a more refined analysis and derive the conditions under which the original inference algorithm in [4] converges to the global minimum. Actually the scenarios analysed in [3] are special cases of Theorem 1 where $\mathcal{L}(\boldsymbol{\alpha}) \to -\infty$.

In addition we observe that a proper scaling of the cost function gives the squared ℓ_2-norm of the reconstruction error. Reconstruction is then equivalent to recovering a support set that minimises the reconstruction error. This principle is effectively the same as the one behind many greedy algorithms such as the OMP [5] and subspace pursuit [6].

4. MODIFIED LIKELIHOOD MAXIMISATION

Theorem 1 suggests certain connections between well-known sparse reconstruction algorithms and SBL. This becomes especially evident when studying the noiseless setting where SBL and sparse signal reconstruction seem to share the same principle. We face the following uncertainty; in [4] selection is based on the value of α_i which maximises the difference $\Delta\mathcal{L}$ in the likelihood function, while in algorithms such as the OMP and SP basis functions are selected based on correlation maximisation and the maximum value of the estimated $\boldsymbol{x}_i$. The following theorem provides the theoretical backing to improve the SBL inference algorithm.

Theorem 2. *Assume the noiseless setting $\boldsymbol{y} = \boldsymbol{\Phi x}$ where $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{\phi}_i^T \boldsymbol{\phi}_i = 1$ for all $1 \leq i \leq n$. Furthermore assume that $t = \max \left| \boldsymbol{\phi}_i^T \boldsymbol{\phi}_j \right|$ for $1 \leq i \neq j \leq n$. Then an algorithm similar to the one in [4] based on one of the following criteria recovers all s-sparse signals exactly given the sufficient condition $t < 0.375/s$; (a) the maximum $\sigma^2 \Delta\mathcal{L}$, (b) the maximum x_i or (c) the minimum α_i.*

Proof: Consider the index set $\mathcal{I}$ as defined earlier in Theorem 1. Then:

$$\begin{aligned}\boldsymbol{C}_{-i}^{-1} &= \left(\boldsymbol{I} + \boldsymbol{\Phi}_{\mathcal{D}}\boldsymbol{A}_{\mathcal{D}}^{-1}\boldsymbol{\Phi}_{\mathcal{D}}^{T}\right)^{-1} \\ &= \sigma^{-2}\left(\boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{D}}(\sigma^{2}\boldsymbol{A}_{\mathcal{D}} + \boldsymbol{\Phi}_{\mathcal{D}}^{T}\boldsymbol{\Phi}_{\mathcal{D}})^{-1}\boldsymbol{\Phi}_{\mathcal{D}}^{T}\right)\end{aligned}$$

where $\mathcal{D} = \mathcal{I} - \{i\}$. By properly scaling as $\sigma^2 \to 0$ then:

$$\lim_{\sigma^2\to 0} \sigma^2 \boldsymbol{C}_{-i}^{-1} = \boldsymbol{I} - \boldsymbol{\Phi}_{\mathcal{D}}\boldsymbol{\Phi}_{\mathcal{D}}^{\dagger}$$

Hence in the limit of zero noise variance σ^2 the optimal value for each hyper-parameter is given by:

$$\begin{aligned}\alpha_i &= \frac{\sigma^4 s_i^2}{\sigma^4 q_i^2 - \sigma^4 s_i} = \frac{\bar{s}_i^2}{\bar{q}_i^2 - \sigma^2 \bar{s}_i} \\ &= \frac{(\boldsymbol{\phi}_i^T(\boldsymbol{\phi}_i - \boldsymbol{\Phi}_{\mathcal{D}}\boldsymbol{\Phi}_{\mathcal{D}}^{\dagger}\boldsymbol{\phi}_i))^2}{(\boldsymbol{\phi}_i^T(\boldsymbol{y} - \boldsymbol{\Phi}_{\mathcal{D}}\boldsymbol{\Phi}_{\mathcal{D}}^{\dagger}\boldsymbol{y}))^2 + \sigma^2(\boldsymbol{\phi}_i - \boldsymbol{\Phi}_{\mathcal{D}}\boldsymbol{\Phi}_{\mathcal{D}}^{\dagger}\boldsymbol{\phi}_i))} \\ \lim_{\sigma^2\to 0} \alpha_i &= \frac{(\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i,r})^2}{(\boldsymbol{\phi}_i^T\boldsymbol{y}_{i,r})^2} = \frac{\bar{s}_i^2}{\bar{q}_i^2}, \end{aligned} \tag{4}$$

where $\boldsymbol{\phi}_{i,r}$ denotes the residual vector from the projection of $\boldsymbol{\phi}_i$ on the span of $\boldsymbol{\Phi}_{\mathcal{D}}$. The same holds for $\boldsymbol{y}_{i,r}$.

By the means of Equation (4) it becomes evident that recovering all s-sparse signals $\boldsymbol{x}$ in the noiseless case, is the same as recovering the support set for which α_i is minimised. This is the same as maximising the denominator in (4) since the numerator is bounded:

$$1 - st \leq |\boldsymbol{\phi}_i^T\boldsymbol{\phi}_{i,r}| \leq 1$$

where it is assumed that $t \geq |\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j|,\ \forall\ 1 \leq i \neq j \leq n$. Let

$$j^* = \arg\min_j\ \alpha_j \text{ so that } j^* \in \mathcal{I}.$$

For any $j \notin \mathcal{I}$:

$$|\boldsymbol{\phi}_j^T\boldsymbol{y}| \leq t \parallel \boldsymbol{x} \parallel_1 \leq t\sqrt{s} \parallel \boldsymbol{x} \parallel_2$$

The above is a direct result of Hölder's inequality. Similarly for any $j \in \mathcal{I}$:

$$\begin{aligned}|\boldsymbol{\phi}_j^T y| &\geq |x_j| - \left|\sum_{i\neq j} x_i \boldsymbol{\phi}_i^T\boldsymbol{\phi}_j\right| \geq |x_j| - t\sum_{i\neq j}|x_i| \\ &\geq |x_j| - t\sqrt{s} \parallel \boldsymbol{x} \parallel_2 \geq \left(\frac{1}{\sqrt{s}} - t\sqrt{s}\right)\|\boldsymbol{x}\|_2\end{aligned}$$

since: $\parallel \boldsymbol{x}_{j\in\mathcal{I}} \parallel_\infty \geq \frac{1}{\sqrt{K}} \parallel \boldsymbol{x} \parallel_2$. By combining the results above, for $j \notin \mathcal{I}$:

$$\alpha_i \geq \frac{1 - st}{t\sqrt{s}\,\|\boldsymbol{x}\|_2}$$

and for $j \in \mathcal{I}$:

$$\alpha_i \leq \frac{1}{\left(\frac{1}{\sqrt{s}} - t\sqrt{s}\right)\|\boldsymbol{x}\|_2}$$

The following must hold in order for $j^* \in \mathcal{I}$:

$$\begin{gathered}\frac{1}{\left(\frac{1}{\sqrt{s}} - t\sqrt{s}\right)\|\boldsymbol{x}\|_2} < \frac{1 - st}{t\sqrt{s}\,\|\boldsymbol{x}\|_2} \\ s^2t^2 - 3st + 1 < 0\end{gathered}$$

By solving the inequality above we get, $t < \frac{3-\sqrt{5}}{2s} \approx \frac{0.375}{s}$ which proves the first part of the theorem.

It is imperative that we derive the correct expressions for the corresponding values of x_i and the change in the cost function $\Delta\mathcal{L}$ for when $\sigma^2 = 0$:

$$x_i = \frac{\bar{q}_i}{\bar{s}_i^2}, \quad \sigma^2\Delta\mathcal{L}_i = \frac{\bar{q}_i}{\bar{s}_i} \tag{5}$$

where subscript i again denotes the association with one particular component. By comparing the expressions in (5) and (4) the proof for second part of the theorem follows.

According to Theorem 2 it is possible to re-design the inference algorithm to achieve better performance when the noise variance approaches zero but at the same time retain the flexibility of the Bayesian approach when noise variance is unknown but somehow estimated. Variants of the inference algorithm can be constructed by employing a different selection criterion according to Theorem 2. In Algorithm 1 a variant is presented in which the maximum value of $|\mathbf{x}_i|$ is used as a selection criterion. The rest of the variants are not presented here for brevity.

Algorithm 1 FMLM-$\boldsymbol{x}_i$

```
Input: $\boldsymbol{\Phi}, \boldsymbol{y}, \sigma^2$
Initialise:
  - $\hat{T} = \{$index $i \in [1, n]$ for maximum $|\boldsymbol{\phi}_i^T\boldsymbol{y}|\}$.
Iteration:
  - Calculate values of $\alpha_i$ and $[\boldsymbol{\mu}_x]_i$ for $i \in [1, n] \setminus \hat{T}$.
  - $T' = \hat{T} \cup \{$index $i$ corresponding to the maximum value
  of $[\boldsymbol{\mu}_{\boldsymbol{x}}]_i$ for $i \notin \hat{T}\}$.
  - Calculate values $\alpha_i$ for $i \in T'$.
  - $\tilde{T} = \{i \in T' : 0 < \alpha_i < +\infty\}$.
  - If $|\bar{\mathcal{L}}_{\tilde{T}} - \bar{\mathcal{L}}_{\hat{T}}| = 0$ then compute $\sigma^{-2}\boldsymbol{\Sigma}_x$, $\boldsymbol{\mu}_x$ for $\tilde{T}$ and
  quit. Set $\hat{T} = \tilde{T}$ and continue otherwise.
Output:
  - Estimated support set $\tilde{T}$ and sparse signal $\tilde{\boldsymbol{x}}$ with $|\tilde{T}|$
  non-zero components, $\tilde{\boldsymbol{x}}_{\tilde{T}} = \boldsymbol{\mu}_x$.
  - Estimated covariance matrix $\sigma^{-2}\boldsymbol{\Sigma}_x$.
```

5. BAYESIAN SUBSPACE PURSUIT

Given the results from Theorem 2 and the subsequent algorithm re-design in Algorithm 1 we move forward in completely altering the inference procedure with ideas from SP [6]. As an intermediate result we point out that by using $\theta_i = \bar{q}_i$ as the selection criterion the inference algorithm achieves the same performance guarantees as the OMP since both techniques become equivalent. The algorithm, termed henceforth Bayesian Subspace Pursuit is described in Algorithm 2.

Algorithm 2 Bayesian Subspace Pursuit

```
Input: $\mathbf{\Phi}, \boldsymbol{y}, \sigma^2$
Initialise:
  - $\hat{T} = \{\text{index } i \in [1, n] \text{ for minimum } \alpha_i = \frac{1}{|\boldsymbol{\phi}_i^T \boldsymbol{y}|}\}$.
Iteration:
  - Store $\alpha_{max} = \arg\max_{i \in \hat{T}} |\alpha_i|$.
  - Calculate values $\alpha_i$ and $\theta_i = \bar{q}_i^2 - \bar{s}_i$ for $i \in [1, n]$.
  - Calculate values $t_{\theta_i > 0} = |\{i \in [1, n] : \theta_i > 0\}|$ and
    $t_{\alpha_i \leq a_{max}} = |\{i \in [1, n] : |\alpha_i| \leq a_{max}\}|$.
  - If $t_{\theta_i > 0} = 0$ then $s = t_{\alpha_i \leq a_{max}} + 1$ else
    $s = t_{\theta_i > 0} + t_{\alpha_i \leq a_{max}}$.
  - $T' = \hat{T} \cup \{$indices corresponding to $s$ smallest values of
    $\alpha_i$ for $i \in [1, n]\}$.
  - Compute $\sigma^{-2}\mathbf{\Sigma}_x$ and $\boldsymbol{\mu}_x$ for $T'$.
  - $\tilde{T} = \{$indices corresponding to $s$ largest non-zero values
    of $|\boldsymbol{\mu}_x|$ for which $0 < \alpha_i < +\infty\}$.
  - If $|\bar{\mathcal{L}}_{\tilde{T}} - \bar{\mathcal{L}}_{\hat{T}}| = 0$ then quit. Otherwise set $\hat{T} = \tilde{T}$ and
    continue.
Output:
  - Estimated support set $\tilde{T}$ and sparse signal $\tilde{\boldsymbol{x}}$ with $|\tilde{T}|$
    non-zero components, $\tilde{\boldsymbol{x}}_{\tilde{T}} = \boldsymbol{\mu}_x$.
  - Estimated covariance matrix $\sigma^{-2}\mathbf{\Sigma}_x$ for $\tilde{T}$.
```

Theorem 3. *Assume that the same conditions hold as in Theorem 2. An algorithm similar to the one in [4] based on the less greedy criterion of maximum $\theta_i = \bar{q}_i$, recovers all s-sparse signals exactly given the sufficient condition $t < 0.5/s$. The algorithm presented in Algorithm 2 recovers all s-sparse signals exactly if matrix $\mathbf{\Phi}$ satisfies the RIP with parameter $\delta_{3s} < 0.205$.*

Proof: The first part of the theorem can be easily proven by adopting the same procedure as in Theorem 2 by replacing α_i with $\bar{q}_i$ as the selection criterion. We arrive at $t < 0.5$ which is equivalent to the mutual coherence restriction for the OMP [5]. For the rest of the theorem proof is heavily based on results from [6]. More specifically by applying Theorems 3 and 4 from [6] one arrives at:

$$\rho = \frac{\| \boldsymbol{y}_{r,\tilde{T}} \|_2}{\| \boldsymbol{y}_{r,\hat{T}} \|_2} \leq \frac{1 + \delta_{3s}}{1 - 2\delta_{3s}} \frac{\sqrt{10\delta_{3s}}}{1 - \delta_{3s}}$$

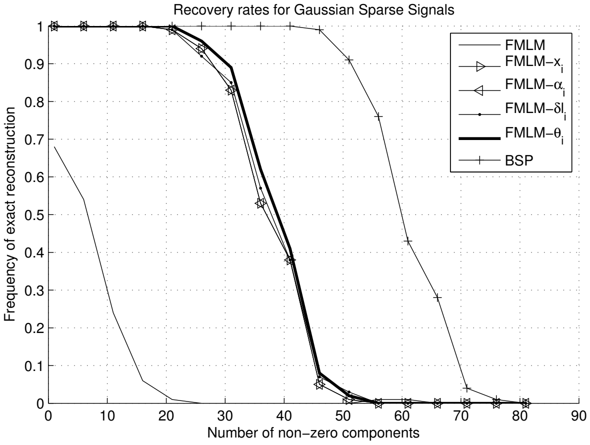

Fig. 1. Exact reconstruction rates for $m = 128, n = 256$

By requiring that $\rho < 1$ and after some basic calculations one arrives at $\delta_{3s} < 0.205$.

6. EMPIRICAL RESULTS

The algorithms under comparison are the FMLM algorithm as originally presented in [4], the variants based on the scaled quantities; FMLM-$\boldsymbol{x}_i$, FMLM-$\boldsymbol{\alpha}_i$, FMLM-$\Delta\mathcal{L}_i$, FMLM-$\boldsymbol{\theta}_i$ and the BSP. The experiment is as follows:

1. Generate $\mathbf{\Phi} \in \mathbb{R}^{128 \times 256}$ with i.i.d entries from $\mathcal{N}\left(0, \frac{1}{m}\right)$.
2. Generate T uniformly at random so that $|T| = K$.
3. Choose values for $\boldsymbol{x}_T$ from $\mathcal{N}(0, 1)$.
4. Compute $\boldsymbol{y} = \mathbf{\Phi}\boldsymbol{x}$ and apply a reconstruction algorithms. Compare estimate $\hat{\boldsymbol{x}}$ to $\boldsymbol{x}$.
5. Repeat experiment for increasing values of K and for 100 realisations.

The results from this procedure are depicted in Figure 1. At first we can see that the original FMLM performs poorly when $\sigma^2 = 0$ due to the improperly scaled cost function. The three variants of FMLM based on Theorem 2 perform according to the theorem. We observe an increase in the performance for FMLM-$\boldsymbol{\theta}_i$, a consequence of altering the selection criterion to $\theta_i = \bar{q}_i$. Even though changing the criterion gives theoretically better performance as Theorem 3 suggests, empirically this gain is not great. By redesigning the inference algorithm based on ideas from the SP we are able to achieve far better performance, as the curve for the BSP algorithm shows.

7. CONCLUSION

We have presented a theoretical analysis of the algorithms in [4]. It became clear that the algorithms can be redesigned as to achieve greater performance guarantees. The theoretical ground has been provided to justify the proposed approach. We also provide simulation results that experimentally verify the improved inference algorithms.

8. REFERENCES

[1] Michael E. Tipping, "The relevance vector machine," 2000.

[2] Shihao Ji, Ya Xue, and L. Carin, "Bayesian compressive sensing," *IEEE Trans. Signal Process.*, vol. 56, no. 6, pp. 2346 –2356, june 2008.

[3] D.P. Wipf and B.D. Rao, "Sparse Bayesian learning for basis selection," *IEEE Trans. Signal Process.*, vol. 52, no. 8, pp. 2153 – 2164, aug. 2004.

[4] M.E. Tipping, A.C. Faul, et al., "Fast marginal likelihood maximisation for sparse Bayesian models," in *International workshop on artificial intelligence and statistics*. Jan, 2003, vol. 1.

[5] J.A. Tropp, "Greed is good: algorithmic results for sparse approximation," *IEEE Trans. Inform. Theory*, vol. 50, no. 10, pp. 2231 – 2242, oct. 2004.

[6] Wei Dai and Olgica Milenkovic, "Subspace pursuit for compressive sensing signal reconstruction," *IEEE Trans. Inform. Theory*, vol. 55, pp. 2230–2249, 2009.