Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

IEEE Trans Signal Process. Author manuscript; available in PMC 2021 June 28.

Published in final edited form as:

IEEE Trans Signal Process. 2018 June 15; 66(12): 3124–3139. doi:10.1109/tsp.2018.2824286.

Rectified Gaussian Scale Mixtures and the Sparse Non-Negative Least Squares Problem

Alican Nalci [Student Member, IEEE],
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Igor Fedorov [Student Member, IEEE],
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Maher Al-Shoukairi [Student Member, IEEE],
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Thomas T. Liu [Member, IEEE],
Departments of Radiology, Psychiatry and Bioengineering, and UCSD Center for Functional MRI, University of California, San Diego, 9500 Gilman Drive,CAN La Jolla, CA 92093, USA

Bhaskar D. Rao [Fellow, IEEE]
Department of Electrical and Computer Engineering, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract

In this paper, we develop a Bayesian evidence maximization framework to solve the sparse non-negative least squares problem (S-NNLS). We introduce a family of probability densities referred to as the Rectified Gaussian Scale Mixture (R-GSM), to model the sparsity enforcing prior distribution for the signal of interest. The R-GSM prior encompasses a variety of heavy-tailed distributions such as the rectified Laplacian and rectified Student-t distributions with a proper choice of the mixing density. We utilize the hierarchical representation induced by the R-GSM prior and develop an evidence maximization framework based on the Expectation-Maximization (EM) algorithm. Using the EM-based method, we estimate the hyper-parameters and obtain a point estimate for the solution of interest. We refer to this proposed method as rectified Sparse Bayesian Learning (R-SBL). We provide four EM-based R-SBL variants that offer a range of options to trade-off computational complexity to the quality of the E-step computation. These methods include the Markov Chain Monte Carlo EM, linear minimum mean square estimation, approximate message passing and a diagonal approximation. Using numerical experiments, we

Personal use is permitted, but republication/redistribution requires IEEE permission.

Correspondence: analci@ucsd.edu.

[3]Practically it was found that setting $\tau_{x_i} = 0$ when $\hat{x}_i < 0$ increases the chances of the algorithm getting stuck at a local minimum. Instead, we set $\tau_{x_i} = \frac{\tau_{r_i}\gamma_i}{\tau_{r_i}+\gamma_i} = \nu_i$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

show that the proposed R-SBL method outperforms existing S-NNLS solvers in terms of both signal and support recovery, and is very robust against the structure of the design matrix.

Index Terms

Non-negative Least Squares; Sparse Bayesian learning; Sparse Signal Recovery; Rectified Gaussian Scale Mixtures

I. Introduction

This work considers the following signal model

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

where the solution vector $\boldsymbol{x} \in \mathbb{R}_+^M$ is assumed to be non-negative, the matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N \times M}$ is fixed and obtained from the physics of the underlying problem, $\boldsymbol{y} \in \mathbb{R}^N$ is the measurement, and $\boldsymbol{v}$ is the additive noise, modeled as a zero mean Gaussian with uncorrelated entries e.g. $\boldsymbol{v} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.

Recovering $\boldsymbol{x}$ using the signal model in Eq. (1) is known as solving the non-negative least squares (NNLS) problem. NNLS has a rich history in the context of methods for solving systems of linear equations [1], density estimation [2] and non-negative matrix factorization (NMF) [3], [4], [5], [6]. NNLS is also widely used in applications such as text mining [7], image hashing [8], speech enhancement [9], spectral decomposition [10], magnetic resonance chemical shift imaging [11] and impulse response estimation [12]. The maximum-likelihood solution for the model in Eq. (1) is given by

$$\underset{\boldsymbol{x} \geq \boldsymbol{0}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2 . \tag{2}$$

In many applications, $N < M$ and Eq. (1) is under-determined, meaning that there are more unknowns than equations and a unique solution to Eq. (2) may not exist. Recovering a unique solution is possible, if more information is known *a-priori* about the solution vector. For example, a useful assumption is that the solution vector is *sparse* and contains only a few non-zero elements [13], [14], [15]. In this case, the sparsest solution (e.g. assuming a noiseless scenario) can be recovered by modifying Eq. (2) to

$$\underset{\boldsymbol{x} \geq \boldsymbol{0},\, \boldsymbol{y} = \boldsymbol{\Phi x}}{\text{minimize}} \quad \|\boldsymbol{x}\|_0, \tag{3}$$

where $\|.\|_0$ is the ℓ_0 pseudo-norm, which counts the non-zero elements in $\boldsymbol{x}$. The number of non-zero elements in $\boldsymbol{x}$ is also referred to as the cardinality of the solution. Then, the design objective in Eq. (3) is to minimize the number of non-zero elements in $\boldsymbol{x}$, while satisfying the optimization constraints. We refer to Eq. (3) as the sparse NNLS (S-NNLS) problem.

The S-NNLS problem is becoming increasingly popular in certain applications, where the non-negative solution has to be recovered from a limited number of measurements. For

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

example, in [16] an S-NNLS recovery method was applied to magnetic resonance imaging (MRI) data to reconstruct narrow fiber-crossings from a limited number of acquisitions. In [17], another method was used to uncover regulatory networks from micro-array mRNA expression profiles from breast cancer data. In [18], [19], an S-NNLS method was applied to functional MRI data to estimate sparsely repeating spatio-temporal activation patterns in the human brain. S-NNLS solvers are also used in applied mathematics for creating dictionaries for sparse representations, such as sparse NMF and non-negative K-SVD [3], [20].

Directly solving Eq. (3) is not tractable, since the ℓ_0 penalty is not convex and the problem is NP-hard [21], [22]. Therefore, 'greedy' algorithms have been proposed to approximate the solution [23], [24], [25], [26], [27]. An example is the class of algorithms known as Orthogonal Matching Pursuit (OMP) [23], [28], which greedily select snon-zero elements of $\boldsymbol{x}$. In order to adapt OMP to the S-NNLS problem, the criterion by which a new non-zero element of $\boldsymbol{x}$ is selected is modified to select the one having the largest *positive* value [27]. Another approach in this class of algorithms finds an $\boldsymbol{x}$, such that $\|\boldsymbol{y} - \Phi\, \boldsymbol{x}\|_2 \leq \boldsymbol{\varepsilon}$ and $\boldsymbol{x} \geq 0$ using the active-set Lawson-Hanson algorithm [1] and then prunes $\boldsymbol{x}$ until $\|\boldsymbol{x}\|_0 \leq K$, where K is a pre-specified desired cardinality [3].

Greedy algorithms are computationally attractive, however they may lead to sub-optimal solutions. Therefore, convex relaxations of the ℓ_0 penalty have been proposed [22], [29], [30], [31], [32]. One simple alternative replaces the ℓ_0 with an ℓ_1 norm and reformulate the problem in Eq. (3) as

$$\underset{\boldsymbol{x} \geq \mathbf{0}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2 + \lambda \|\boldsymbol{x}\|_1, \tag{4}$$

where $\lambda > 0$ is a suitably chosen regularization parameter to account for the measurement noise. The advantage of the formulation in Eq. (4) is that, it is a convex optimization problem and can be solved by a number of methods [32], [33], [34], [35]. One such approach is to estimate $\boldsymbol{x}$ through projected gradient descent [36].

In fact, the ℓ_1 norm penalty in Eq. (4) can be replaced by any arbitrary sparsity inducing surrogate function $g(\boldsymbol{x})$, thus leading to alternative methods based on solving the following optimization problem

$$\underset{\boldsymbol{x} \geq \mathbf{0}}{\text{minimize}} \quad \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2 + \lambda g(\boldsymbol{x})\,. \tag{5}$$

For example, a surrogate function $g(\boldsymbol{x}) = \sum_{i=1}^{M} \log\,(x_i^2 + \beta)$ has been considered [37], [38], which leads to an iterative reweighted optimization approach.

An alternative view on the S-NNLS problem is to cast the entire problem in a Bayesian framework and consider the maximum a-posteriori (MAP) estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$

$$\boldsymbol{x}_{MAP} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y})\,. \tag{6}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

There is a strong connection between the MAP framework and deterministic formulations like the one in Eq. (5). Recently, it has been shown that formulations of the form in Eq. (5) can be represented by the formulation in Eq. (6) with a proper choice of $p(\boldsymbol{x})$ [39]. For example, considering a separable $p(\boldsymbol{x})$ of the form

$$p(\boldsymbol{x}) = \prod_{i=1}^{M} p(x_i), \tag{7}$$

the ℓ_1 regularization approach in Eq. (4) (e.g. choice of $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ in Eq. (5)) is equivalent to the Bayesian formulation in Eq. (6) with an exponential prior for x_i. In this paper, our emphasis will be on Bayesian approaches for solving Eq. (1).

A. Contributions of the paper

- We introduce a family of non-negative probability densities, referred to as the rectified Gaussian scale mixture (R-GSM) to model non-negative and sparse solutions.
- We discuss how the R-GSM prior encompasses other sparsity promoting non-negative priors, such as the rectified Laplacian and rectified Student-t distributions, through a proper choice of mixing density.
- We detail how the R-GSM prior can be utilized to solve the sparse NNLS problem, using an evidence maximization based estimation procedure that utilizes the expectation-maximization (EM) framework. We refer to this technique as rectified Sparse Bayesian Learning (R-SBL).
- We provide four alternative EM-based approaches, that offer a range of options to trade-off computational complexity to the quality of the E-step computation, including the Markov Chain Monte Carlo EM, linear minimum mean square estimation, approximate message passing and a diagonal approximation.
- We use extensive empirical results to show the robustness and superiority of the R-GSM priors and R-SBL algorithm for the S-NNLS problem. Especially, under various i.i.d. and non-i.i.d. settings for the design matrix $\boldsymbol{\Phi}$.

B. Organization of the paper

In Section II, we discuss the advantages of scale mixture priors for $p(\boldsymbol{x})$ and introduce the Rectified Gaussian Scale Mixture (R-GSM) prior. In Section III, we define the Type I and Type II Bayesian approaches to solve the S-NNLS problem and introduce the R-SBL framework with the R-GSM prior. We provide details of an evidence maximization based estimation procedure in Section III-B. We present empirical results comparing the proposed R-SBL algorithm to existing baseline methods in Section V.

II. Rectified Gaussian Scale Mixtures

We assume separable priors of the form in Eq. (7) and focus on the choice of $p(x_i)$. The choice of prior plays a central role in the Bayesian inference [40], [41], [42]. For the S-

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

NNLS problem, our prior must be sparsity inducing and satisfy the non-negativity constraints. Consequently, we consider the hierarchical scale mixture prior

$$p(x_i) = \int_0^\infty p(x_i|\gamma_i)p(\gamma_i)d\gamma_i\,. \tag{8}$$

Scale mixture priors were first considered in the form of Gaussian Scale Mixtures (GSM) where $p(x_i|\gamma_i) = \mathcal{N}(x_i;\, 0,\, \gamma_i)$ [43]. Super-gaussian densities are suitable priors for promoting sparsity [40], [44] and most of those priors can be represented in the form shown in Eq. (8) with a proper choice of $p(\gamma_i)$ [45], [46], [47], [48], [49]. This has made scale mixture based priors valuable for the general sparse signal recovery problem. Another advantage of the scale mixture prior is that, it establishes a Markovian structure of the form

$$\boldsymbol{\gamma} \rightarrow \boldsymbol{x} \rightarrow \boldsymbol{y}, \tag{9}$$

where inference can be performed in the $\boldsymbol{x}$ domain (referred to as Type I) and in the $\boldsymbol{\gamma}$ domain (Type II). Experiment results in the standard sparse signal recovery problem show that performing inference in the $\boldsymbol{\gamma}$ domain consistently achieves superior performance [39], [40], [50], [51].

The Type II procedure involves finding a maximum-likelihood (ML) estimate of $\boldsymbol{\gamma}$ using evidence maximization and approximating the posterior $p(\boldsymbol{x}\,|\,\boldsymbol{y})$ by $p(\boldsymbol{x}\,|\,\boldsymbol{y},\, \boldsymbol{\gamma}_{\mathrm{ML}})$. The performance gains can be understood by noting that $\boldsymbol{\gamma}$ is deeper in the Markovian chain than $\boldsymbol{x}$, so the influence of errors in performing inference in the $\boldsymbol{\gamma}$ domain may be diminished [39], [50]. Also, $\boldsymbol{\gamma}$ is close enough to $\boldsymbol{y}$ in the Markovian chain such that meaningful inference about $\boldsymbol{\gamma}$ can still be performed, mitigating the problem of local minima that is more prevalent when seeking a Type I estimate of $\boldsymbol{x}$ [50].

Although priors of the form shown in Eq. (8) have been used in the compressed sensing literature (where the signal model is identical to Eq. (1) without the non-negativity constraint) [39], [52], [53], such priors have not been extended to solve the S-NNLS problem. Considering the findings that the scale mixture prior has been useful for the development of sparse signal recovery algorithms [39], [50], [54], we propose a R-GSM prior for the S-NNLS problem, where $p(x_i|\gamma_i)$ in Eq. (8) is the rectified Gaussian (RG) distribution. We refer to the proposed Type II inference framework as R-SBL.

The RG distribution is defined as

$$\mathcal{N}^R(x;\mu,\gamma) = \sqrt{\frac{2}{\pi\gamma}}\frac{e^{-\frac{(x-\mu)^2}{2\gamma}}u(x)}{\operatorname{erfc}\left(-\frac{\mu}{\sqrt{2\gamma}}\right)}, \tag{10}$$

where μ is the location parameter (e.g. is not the mean), γ is the scale parameter, $u(x)$ is the unit step function, and $\operatorname{erfc}(x)$ is the complementary error function.[1]

[1] $\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}\,dt$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

As noted in previous works [55], [56], closed form inference computations using a multivariate RG distribution are tractable only if the location parameter is zero (e.g. by effectively getting rid of the erfc(.) term). Although a non-zero μ could provide a richer class of priors, possibly to model approximately sparse or non-sparse solutions, considering the tractability issues and the potential overfitting problems (twice as many parameters), we focus on the R-GSM priors with $\mu = 0$ to promote *sparse* non-negative solutions. It is a pragmatic choice and adequate for the problem at hand.

When $\mu = 0$, the RG density simplifies to

$$\mathcal{N}^R(x; 0, \gamma) = \sqrt{\frac{2}{\pi\gamma}} e^{-\frac{x^2}{2\gamma}} u(x)\,. \tag{11}$$

Thus, the R-GSM prior introduced in this work have the form

$$p(x) = \int_0^\infty \mathcal{N}^R(x; 0, \gamma) p(\gamma) d\gamma\,. \tag{12}$$

Different choices of $p(\gamma)$ lead to different choices of priors and some examples are presented below.

A. R-GSM representation of sparse priors

We can utilize the proposed R-GSM framework to obtain other sparse non-negative priors. For instance, consider the rectified Laplace prior $p(x) = \lambda e^{-\lambda x} u(x)$. By using an exponential prior for $p(\gamma) = \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \gamma}{2}} u(\gamma)$, we can express $p(x)$ in the R-GSM framework [57] as

$$p(x) = 2u(x) \int_0^\infty \mathcal{N}(x|0, \gamma) \frac{\lambda^2}{2} e^{-\frac{\lambda^2 \gamma}{2}} u(\gamma) d\gamma \tag{13}$$

$$= \lambda e^{-\lambda x} u(x)\,. \tag{14}$$

Similarly, by considering the R-GSM with $p(\gamma)$ given by the Gamma(a, b) distribution, we obtain a rectified Student-t distribution for $p(x)$, and Eq. (8) simplifies as in [40] to

$$p(x) = 2u(x) \int_0^\infty \mathcal{N}(x|0, \gamma) \frac{\gamma^{a-1} e^{\frac{-\gamma}{b}}}{a^b \Gamma(\mathrm{a})} d\gamma \tag{15}$$

$$= \frac{2b^a \Gamma(a + \frac{1}{2})}{(2\pi)^{\frac{1}{2}} \Gamma(a)} \left(b + \frac{x^2}{2}\right)^{-(a+\frac{1}{2})} u(x), \tag{16}$$

where Γ is defined as $\Gamma(a) = \int_0^\infty t^{a-1} e^{-t}\, dt$. More generally, all of the distributions represented by the GSM have a corresponding rectified version represented by a R-GSM,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

e.g. contaminated Normal and slash, symmetric stable and logistic, hyperbolic, etc. [43], [45], [46], [47], [48], [49].

B. Relation to other Bayesian works

In [55], a modified Gaussian prior was considered for the NNLS problem. The authors used a Gaussian prior of arbitrary mean and variance and performed a non-negative rectification using a 'cut' function. Their goal was to better represent *non-sparse* signals by avoiding the selection of $\mu = 0$, as we consider in our work. Our R-GSM prior substantially differs from this work as we consider a *mixture* of zero-location RG distributions for the prior, as opposed to a single Gaussian density with the 'cut' rectification. Our design objective is to induce sparsity by using a hierarchical hyper-parameter $\boldsymbol{\gamma}$.

In [58], a non-negative generalized approximate message passing (GAMP) approximation was proposed, using a Bernoulli non-negative Gaussian mixture prior of arbitrary location and scale parameters. This extended the prior given in [55] but used a fixed number of mixture components e.g. $L = 3$. The sparsity was enforced by using a Dirac delta function and an additional sparsity rate λ that would 'favor' the Dirac function and attenuate other mixture components, simultaneously. The authors had to infer a bulk of parameters including the scale, location, and mixture weights as well as the sparsity rate simultaneously. Our R-SBL approach differs from [58] as we only consider a sparsity inducing hyper-parameter $\boldsymbol{\gamma}$, and our mixture components are strictly located at zero. Our approach simplifies the overall inference procedure and the problem formulation. We also consider an infinite number of mixture components, as opposed to considering a fixed number of components.

Finally, we consider a more general class of priors than the existing methods since the R-GSM prior is based on an arbitrary mixing density $p(\boldsymbol{\gamma})$. As indicated in Section II-A, different selections of $p(\boldsymbol{\gamma})$ lead to more flexible and a more generalized priors for the spase solution.

III. Bayesian Inference with Scale Mixture Prior

We detail the Type I and Type II methods for solving the S-NNLS problem with an R-GSM prior. Though this paper is dedicated to Type II estimation because of its superior performance in sparse signal recovery problems [39], [50], we briefly introduce Type I in the following section for the sake of completeness.

A. Type I estimation

Using Type I to solve the S-NNLS problem translates into calculating the MAP estimate of $\boldsymbol{x}$ given $\boldsymbol{y}$

$$\arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2 - \lambda \sum_{i=1}^{M} \ln\, p(x_i)\,. \tag{17}$$

Some of the ℓ_0 relaxation methods described in Section I can be derived from a Type I perspective. For instance, by choosing an exponential prior for $p(x_i)$, Eq. (17) reduces to the

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ℓ_1 regularization approach in Eq. (4) with the interpretation of λ as being determined by the parameters of the prior and the noise variance. Similarly, by choosing a Gamma prior for $p(x_i)$, Eq. (17) reduces to

$$\arg\min_{\boldsymbol{x}} \|\boldsymbol{y}-\boldsymbol{\Phi x}\|_2^2+\lambda\sum_{i=1}^{M}\ln\left(b+\frac{x_i^2}{2}\right), \tag{18}$$

which leads to the reweighted ℓ_2 approach to the S-NNLS problem described in [37], [38]. A unified Type I approach for the R-GSM prior can be readily derived using the approaches as discussed in [39], [45].

B. Type II estimation

The Type II framework involves finding a ML estimate of $\boldsymbol{\gamma}$ using evidence maximization and approximating the posterior $p(\boldsymbol{x}\,|\,\boldsymbol{y})$ by $p(\boldsymbol{x}\,|\,\boldsymbol{y},\,\boldsymbol{\gamma}_{\mathrm{ML}})$. Then, appropriate point estimates and the solution $\boldsymbol{x}$ can be estimated. We refer to this approach as the Rectified Sparse Bayesian Learning (R-SBL).

Several strategies exist for estimating $\boldsymbol{\gamma}$. The first strategy considers the problem of forming a ML estimate of $\boldsymbol{\gamma}$ given $\boldsymbol{y}$ [39], [40], [59], [60]. In our case, $p(\boldsymbol{\gamma}\,|\,\boldsymbol{y})$ does not admit a closed form expression, making this approach difficult. The second strategy, which is investigated in this paper, aims to estimate $\boldsymbol{\gamma}$ by utilizing an EM algrotihm [39], [52], [60]. In the EM approach, we treat $(\boldsymbol{x},\,\boldsymbol{y},\,\boldsymbol{\gamma})$ as the complete data and $\boldsymbol{x}$ as the hidden variable. Utilizing the current estimate $\boldsymbol{\gamma}^t$, where t refers to the iteration index, the expectation step (E-step) involves determining the expectation of the log-likelihood, $Q(\boldsymbol{\gamma},\,\boldsymbol{\gamma}^t)$ given by

$$Q(\boldsymbol{\gamma},\boldsymbol{\gamma}^t)=E_{\boldsymbol{x}\,|\,\boldsymbol{y};\,\boldsymbol{\gamma}^t}[\ln\,p(\boldsymbol{y}|\boldsymbol{x})+\ln\,p(\boldsymbol{x}\,|\,\boldsymbol{\gamma})+\ln\,p(\boldsymbol{\gamma})] \tag{19}$$

$$\doteq\sum_{i=1}^{M}E_{\boldsymbol{x}\,|\,\boldsymbol{y};\,\boldsymbol{\gamma}^t}\left[-\frac{1}{2}\ln\gamma_i-\frac{x_i^2}{2\gamma_i}+\ln\,p(\gamma_i)\right], \tag{20}$$

where $\doteq$ indicates that constant terms, and terms that do not depend on $\boldsymbol{\gamma}$ have been dropped since they do not affect the M-step. For simplicity, we assume a non-informative prior on $\boldsymbol{\gamma}$ [40]. In the M-step, we maximize $Q(\boldsymbol{\gamma},\,\boldsymbol{\gamma}^t)$ with respect to $\boldsymbol{\gamma}$ by taking the derivative and setting it equal to zero, which yields the update rule

$$\gamma_i^{t+1}=E_{\boldsymbol{x}\,|\,\boldsymbol{y},\,\boldsymbol{\gamma}^t,\,\sigma^2}[x_i^2]\coloneqq\left\langle x_i^2\right\rangle. \tag{21}$$

To compute $\langle x_i^2\rangle$, we first consider the multivariate posterior density $p(\boldsymbol{x}|\boldsymbol{y},\,\boldsymbol{\gamma},\,\sigma^2)$ which has the form (e.g. see Appendix VII-B for details)

$$p(\boldsymbol{x}\,|\,\boldsymbol{y},\boldsymbol{\gamma})=c(\boldsymbol{y})e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}{2}}u(\boldsymbol{x}), \tag{22}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by [40], [52], [61]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\boldsymbol{\mu} = \boldsymbol{\Gamma}\boldsymbol{\Phi}^T(\sigma^2\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{y} \tag{23}$$

$$\sum = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}\boldsymbol{\Phi}^T(\sigma^2\boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^T)^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}, \tag{24}$$

and $\boldsymbol{\Gamma} = \text{diag}(\boldsymbol{\gamma})$. The posterior in Eq. (III-B) is known as a multivariate RG (or a multivariate truncated normal [62]). The normalizing constant $c(\boldsymbol{y})$ in the multivariate RG does not have a closed form expression. Note that the M-step in Eq. (21) only requires the marginal density. However, the marginals of a multivariate RG are not univariate-RG's, and also do not admit a closed form expression [62], which unfortunately means no immediate closed form expression for the marginal moments.

However, we can approximate the first and the second moments $\langle x_i \rangle$ and $\langle x_i^2 \rangle$ of the multivariate RG posterior. In the following, we propose four different approaches for this purpose that offer a trade-off between computational complexity and theoretical accuracy.

1) Markov Chain Monte Carlo EM (MCMC-EM)—Advances in numerical methods made it possible to sample from complex multivariate distributions [63], [64], [65]. Numerical methods are particularly useful when the first and second order statistics of a posterior density do not have a closed form expression. In this case, the E-step can be performed by drawing samples using numerical Markov Chain Monte Carlo (MCMC) and then calculating the sample statistics. This is usually referred as MCMC-EM [66], [67]. First, we consider the Gibbs sampling approach in [68], [69]. We use hat notation to refer to the empirical estimates of various parameters (e.g. $\hat{\boldsymbol{\Sigma}}$).

We use the multivariate truncated normal (TN) definition in [69] and write

$$\text{TN}(\boldsymbol{x}; \hat{\boldsymbol{\mu}}, \widehat{\sum}, \boldsymbol{R}, \boldsymbol{\alpha}_L, \boldsymbol{\alpha}_U) = \tag{25}$$

$$\left(c_{tn} e^{-\frac{(\boldsymbol{x}-\hat{\boldsymbol{\mu}})^T \hat{\sum}^{-1} (\boldsymbol{x}-\hat{\boldsymbol{\mu}})}{2}} \right) \mathbf{1}_{\boldsymbol{\alpha}_L \leq \boldsymbol{R}\boldsymbol{w} \leq \boldsymbol{\alpha}_U}, \tag{26}$$

where $\mathbf{1}_{(\cdot)}$ is the indicator function and c_{tn} is the normalizing constant for the density. In the case of a multivariate rectified Gaussian, the truncation bounds are $\boldsymbol{\alpha}_L = \mathbf{0}$ and $\boldsymbol{\alpha}_U = \infty$, and $\mathbf{R} = \mathbf{I}$. By introducing the transformation, $\boldsymbol{w} = \hat{\boldsymbol{L}}^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}})$ where $\hat{\boldsymbol{L}}$ is the lower triangular Cholesky decomposition of $\hat{\boldsymbol{\Sigma}}$, it can be shown that $\boldsymbol{w}$ is $\text{TN}(\boldsymbol{w}; 0, \boldsymbol{I}, \hat{\boldsymbol{L}}, \boldsymbol{\alpha}_L^*, \boldsymbol{\alpha}_U^*)$, with new truncation bounds $\boldsymbol{\alpha}_L^* = \boldsymbol{\alpha}_L - \hat{\boldsymbol{\mu}} = -\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\alpha}_U^* = \boldsymbol{\alpha}_U - \hat{\boldsymbol{\mu}} = \infty$.

The Gibbs sampler then proceeds by iteratively drawing samples from the conditional distribution $p(w_i | \boldsymbol{y}, \hat{\boldsymbol{\gamma}}, \sigma^2, \boldsymbol{w}_{-i})$, where $\boldsymbol{w}_{-i}$ refers to the vector containing all but the ith element of $\boldsymbol{w}$. Given a set of samples drawn from $\boldsymbol{w}$, we can generate samples from the original distribution of interest by inverting the transformation: $\{\boldsymbol{x}^n\}_{n=1}^N = \{\hat{\boldsymbol{L}}\boldsymbol{w}^n + \hat{\boldsymbol{\mu}}\}_{n=1}^N$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Then, the first and second empirical moments can be calculated from the drawn samples using

$$\left\langle x_i \right\rangle \approx \frac{1}{N} \sum_{n=1}^{N} (x_i^n), \tag{27}$$

$$\left\langle x_i^2 \right\rangle \approx \frac{1}{N} \sum_{n=1}^{N} (x_i^n)^2, \tag{28}$$

and the EM step can be iterated by updating $\hat{\gamma}_i^{t+1} = \left\langle x_i^2 \right\rangle$.

After convergence, a point estimate for $\boldsymbol{x}$ is needed. The optimal estimator of $\boldsymbol{x}$ in the minimum mean-square-error (MMSE) sense is simply $\hat{\boldsymbol{x}}_{mean} = \langle \hat{x}_i \rangle$. An alternative point estimate is to use $\hat{\boldsymbol{x}}_{mode}$, given by

$$\hat{\boldsymbol{x}}_{mode} = \arg\max_{\boldsymbol{x}} p(\boldsymbol{x} \mid \boldsymbol{y}, \hat{\boldsymbol{\gamma}}, \sigma^2) \tag{29}$$

$$= \arg\min_{\boldsymbol{x} \geq 0} \|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{x}\|_2^2 + \lambda \sum_{i=1}^{M} \frac{x_i^2}{\hat{\gamma}_i}, \tag{30}$$

where Eq. (30) can be solved by any NNLS solver. The estimate $\hat{\boldsymbol{x}}_{mode}$ could be a favorable point estimate because it chooses the peak of $p(\boldsymbol{x} \mid \boldsymbol{y}, \hat{\boldsymbol{\gamma}}, \sigma^2)$, which could be multi-modal and not well-characterized by its mean.

For the sparse recovery problem, we experienced very slow convergence with Gibbs sampling. Convergence was particularly slow for higher problem dimensions and at larger cardinalities. The latter was expected as a sparse solution is harder to recover in those cases. Thus, we resorted to Hamiltonian Monte Carlo (HMC), which is designed specifically for target spaces constrained by linear or quadratic constraints [65]. HMC improves the MCMC mixing performance by using the gradient information of the target distribution [66].

Despite use of the state of the art MCMC techniques, MCMC-EM might still converge to poor local minima solutions and result in sub-optimal performance [70], [71], [72]. Particularly, performance may be even poorer for underdetermined problems. Though MCMC-EM is not thoroughly investigated for sparse recovery problems, here we list four major issues for consideration:

I. Convergence: MCMC-EM based algorithms can get stuck in a local minima depending on the problem dimensions and complexity of the search space. This is true even for well-posed problems [67], [73]. In under-determined problems, the solution set for Eq. (2) may contain many local minima and thus, a good MCMC-EM implementation should try to avoid local minima.

II. Computational Limits: Current MCMC sampling techniques are not optimal for drawing large sample sizes N from high dimensional multivariate posterior

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

densities. Therefore, the number of available samples is often limited by computational constraints [63], [64], [65].

III. Quality of Parameter Estimates: Since the MCMC samples are determined by random sampling at each iteration, the estimates of $\hat{\boldsymbol{\gamma}}$, $\hat{\boldsymbol{\mu}}$, and $\hat{\boldsymbol{\Sigma}}$ depend highly on the quality of the MCMC estimates $\hat{\boldsymbol{x}}$, which in turn affects the quality of next cycle of MCMC samples. This may lead EM algorithm to converge to a sub-optimal solution.

IV. Structure of Empirical $\hat{\boldsymbol{\Sigma}}$: When M is large and the dimensions of the empirical scale matrix are also large, $\hat{\boldsymbol{\Sigma}}$ may be no longer a good numerical estimate [71], [74], [75]. This could happen when the problem is inherently under-determined with $N < M$, and reveals itself as $\hat{\boldsymbol{\Sigma}}$ being close to singular. Therefore, regularization methods for $\hat{\boldsymbol{\Sigma}}$ are often used to alleviate this problem [71], [72].

The scale matrix $\hat{\boldsymbol{\Sigma}}$ has direct control over the search space for MCMC and spurious off-diagonal values tend to increase the number of local-minima. Therefore, to address the issues listed above, we incorporated ideas from prior work to regularize the estimates of $\hat{\boldsymbol{\Sigma}}$:

- As in [71], [72], we assume that $\hat{\boldsymbol{\Sigma}}$ is sparse and we prune its off-diagonal elements, when they drop below a certain threshold T_p. This prevents the spurious off-diagonal values in $\hat{\boldsymbol{\Sigma}}$ from affecting the next cycle of MCMC samples and improves future estimates of $\hat{\boldsymbol{\gamma}}$.
- We incorporate the shrinkage estimation idea presented in [72], [74] and regularize $\hat{\boldsymbol{\Sigma}}$ as a convex sum of the empirical $\hat{\boldsymbol{\Sigma}}$ and a target matrix $\boldsymbol{T}$ such that, $\hat{\boldsymbol{\Sigma}} = \lambda\,\hat{\boldsymbol{\Sigma}} + (1-\lambda)\boldsymbol{T}$. A simple selection is the matrix $\hat{\boldsymbol{\Sigma}}_\beta$ which is $\hat{\boldsymbol{\Sigma}}$ with diagonal elements scaled by a factor β. Though this approach does not guarantee convergence to a global minimum and the solution could still be a local minima or a saddle point solution, we empirically observed better recovery performance.

2) Linear minimum mean square estimation (LMMSE)—The LMMSE approach is motivated by the complexity of the MCMC-EM approach. Examining the parameters being computed, one can interpret them as finding the MMSE estimate of $\boldsymbol{x}$ and the associated MSE. This motivates replacing the MMSE estimate by the simple LMMSE estimate of $\boldsymbol{x}$. The affine LMMSE estimate for $\boldsymbol{x}$ is

$$\hat{\boldsymbol{x}} = \boldsymbol{\mu_x} + \boldsymbol{R_x}\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{R_x}\boldsymbol{\Phi}^T + \sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu_x}), \tag{31}$$

where $\boldsymbol{R_x}$ is the covariance matrix of $\boldsymbol{x}$ and is diagonal. The estimation error covariance matrix is given by [76]

$$\boldsymbol{R_e} = \boldsymbol{R_x} - \boldsymbol{R_x}\boldsymbol{\Phi}^T(\boldsymbol{\Phi}\boldsymbol{R_x}\boldsymbol{\Phi}^T + \sigma^2\boldsymbol{I})^{-1}\boldsymbol{\Phi}\boldsymbol{R_x}\,. \tag{32}$$

To elaborate, in the E-step where $\boldsymbol{\gamma}$ is fixed at $\boldsymbol{\gamma}^t$, the entries of $\boldsymbol{x}$ are independent, and the prior mean and the prior covariance will be equal to the mean and variance of the independent univariate RG distributions with $p(x_i|\gamma_i) = \mathcal{N}^R(0,\ \gamma_i)$. The mean of a univariate rectified Gaussian density with zero location parameter is given by [77]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\mu_{x,i} = \sqrt{\frac{2\gamma_i}{\pi}}, \tag{33}$$

and the variances, which are the diagonal entries of the diagonal matrix $\boldsymbol{R_x}$, are given by

$$R_{x,ii} = \gamma_i(1 - 2/\pi)\,. \tag{34}$$

Using the values of $\boldsymbol{\mu_x}$ and $\boldsymbol{R_x}$ from Eq. (33) and Eq. (34) in Eq. (31) we obtain the LMMSE point estimate for the solution vector. Similarly, the update for $\boldsymbol{\gamma}$ (M-step) is given by

$$\gamma_i = \hat{x}_i^2 + R_{e,ii}\,. \tag{35}$$

This is sufficient to implement the EM algorithm. Upon convergence, the mean point estimate is simply $\hat{\boldsymbol{x}}_{mean} = \hat{\boldsymbol{x}}$, and the mode point estimate can be obtained by utilizing the converged values γ_i in Eq. (30).

3) Generalized approximate message passing (GAMP)—In this section, we present an EM implementation using the generalized approximate message passing algorithm (GAMP) [78]. A different GAMP based non-negative Gaussian mixture approach was used in [58], which uses an i.i.d. Bernoulli non-negative Gaussian mixture prior with a fixed and known mixture order that is independent of M. To overcome the known convergence issues with the type of GAMP algorithm in [58] when a non-i.i.d. design matrix $\boldsymbol{\Phi}$ is used [79], [80], [81], we incorporate the damping technique used in [81] into the proposed R-SBL GAMP algorithm.

GAMP is a low complexity iterative inference algorithm. The low complexity is achieved by applying quadratic and Taylor series approximations to loopy belief propagation. Under the prior $p(\boldsymbol{x})$ and the likelihood $p(\boldsymbol{y}|\boldsymbol{x})$, GAMP can approximate the MMSE estimate when used in the sum-product mode, or it can approximate the MAP estimate when used in its max-sum mode.

The sum-product version of the algorithm computes the mean and variance of approximate marginal posteriors on x_i, where the approximate posteriors are given by

$$p(x_i|r_i;\tau_{r_i}) \propto p(x_i)\mathcal{N}(x_i; r_i, \tau_{r_i}), \tag{36}$$

where r_i approximates a corrupted version (AWGN) of the true x_i as

$$r_i \approx x_i + \bar{r}_i \tag{37}$$

$$\bar{r}_i \sim \mathcal{N}(0, \tau_{r_i})\,. \tag{38}$$

In the large system limit and when the design matrix $\boldsymbol{\Phi}$ is i.i.d sub-Gaussian, the approximation in Eq. (37) was shown to be exact [78], [82]. Therefore, in sum-product mode, the MMSE estimate $\hat{x}_i$ produced by GAMP corresponds to the MMSE estimate of x_i

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

given r_i, and is given by the conditional mean in Eq. (39) below. Where $\boldsymbol{\tau}_{x_i}$ in sum-product GAMP, corresponds to the MMSE associated with the estimate $\hat{x}_i$ and is given in Eq. (40) as the conditional variance of x_i given r_i

$$\hat{x}_i = \mathbb{E}\{x_i | r_i; \tau_{r_i}\} \tag{39}$$

$$\tau_{x_i} = \mathrm{Var}\{x_i | r_i; \tau_{r_i}\}\,. \tag{40}$$

Similarly, in the max-sum version of GAMP, the MAP estimate $\hat{x}_i$ is given using the proximal operator in Eq. (41) as the MAP estimate of x_i given r_i, while $\boldsymbol{\tau}_{x_i}$ in the max-sum GAMP corresponds to the sensitivity of the proximal thresholding and is given by Eq. (42)

$$\hat{x}_i = \mathrm{prox}_{-\ln p(x_i)}(r_i; \tau_{r_i}) \tag{41}$$

$$\tau_{x_i} = \tau_r \mathrm{prox}'_{-\ln p(x_i)}(r_i; \tau_{r_i}) \tag{42}$$

$$\mathrm{prox}_f(\hat{a}, \tau^a) \triangleq \underset{x \in \boldsymbol{R}}{\arg\min}\, f(x) + \frac{1}{2\tau^a}|x - \hat{a}|^2\,. \tag{43}$$

When implementing the EM algorithm, the approximate posterior computed by the sum-product GAMP can be used to efficiently approximate the E-step [83]. Moreover, in the case of max-sum GAMP, even though the algorithm does not provide marginal distributions, in the large system limit and under i.i.d sub-Gaussian $\boldsymbol{\Phi}$ the assumption of r_i being an AWGN corrupted version of the true x_i still holds [82]. Thus, similar to the approach in [58], an extra step can be added to compute marginal distributions using Eq. (36). The computed marginals can be used to approximate the E-step. For the assumed rectified Gaussian scale mixture prior $p(\boldsymbol{x}|\boldsymbol{\gamma})$ the details of finding $\hat{x}_i$ and $\boldsymbol{\tau}_{x_i}$ estimates in both the sum-product and max-sum cases are shown in Appendix VII-A.

Upon the convergence of the GAMP algorithm, the approximate E-step of the EM algorithm is complete, and we can proceed to evaluate the M-step in Eq. (21) as

$$\left\langle x_i^2 \right\rangle = \int_{x_i} x_i^2 p(x | r_i; \tau_{r_i}) = \tau_{x_i} + \hat{x}_i^2\,. \tag{44}$$

The R-SBL GAMP algorithm based on the E and M-steps described above is summarized Table I. We note that for the AWGN case, the output function of the GAMP algorithm used to evaluate $\boldsymbol{s}$ and $\boldsymbol{\tau}_s$ in Table I, is the same for both sum-product and max-sum GAMP [78].

In Table I, all squares, divisions and multiplications are taken element wise. $K_{\max}$ is the maximum allowed number of GAMP iterations, $\varepsilon_{\mathrm{gamp}}$ is the GAMP normalized tolerance parameter, $I_{\max}$ is the maximum allowed number of EM iterations and $\varepsilon_{\mathrm{em}}$ is the EM normalized tolerance parameter.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4) Diagonal approximation (DA)—We know *a-priori* that the posterior in Eq. (III-B) does not admit a closed form expression. However, to implement the EM algorithm, we only need the marginals of the posterior. We first note that, if the scale matrix $\boldsymbol{\Sigma}$ is diagonal then we could evaluate the normalizing constant $c(\boldsymbol{\gamma})$ in closed form, since the multivariate RG posterior can be written as a product of univariate marginals (e.g. see Appendix VII-B for details).

In the diagonal approximation (DA) approach, we resort to approximating the posterior in Eq. (III-B) with a suitable posterior density e.g. $p(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\gamma}) \approx \tilde{p}(\boldsymbol{x}|\boldsymbol{y}, \boldsymbol{\gamma})$, which could be written as a product of independent marginal densities i.e. $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma})$. This approximate posterior density is derived in Appendix VII-B as

$$\tilde{p}(\boldsymbol{x}\,|\,\boldsymbol{y}, \boldsymbol{\gamma}) = \prod_{i=1}^{M} \tilde{p}(x_i\,|\,\boldsymbol{y}, \boldsymbol{\gamma}) \tag{45}$$

$$= \prod_{i=1}^{M} \sqrt{\frac{2}{\pi \sum_{ii}}} \frac{e^{-\frac{(x_i - \mu_i)^2}{2\sum_{ii}}} u(x_i)}{\operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)}, \tag{46}$$

where μ_i is the ith element of $\boldsymbol{\mu}$ and Σ_{ii} is the ith diagonal element of $\boldsymbol{\Sigma}$ obtained using Eqs. (23) and (24). The marginal $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma})$ in Eq. (46) is the univariate RG density defined in Eq. (10). In other words we have $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma}) = \mathcal{N}^R(x_i; \mu_i, \Sigma_{ii})$. Then, the univariate RG marginals are well-characterized by their first and second moments given in [77], with the first moment given as

$$\langle x_i \rangle = \mu_i + \sqrt{\frac{2\sum_{ii}}{\pi}} \frac{e^{-\frac{\mu_i^2}{2\sum_{ii}}}}{\operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)}, \tag{47}$$

and the second moment is given as

$$\langle x_i^2 \rangle = \mu_i^2 + \sum_{ii} + \mu_i \sqrt{\frac{\sum_{ii}}{\pi}} \frac{e^{-\frac{\mu_i^2}{2\sum_{ii}}}}{\operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)}. \tag{48}$$

Note that the moments of $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma})$ are approximations to the moments of the exact marginals which do not admit closed form. However, we can perform EM using the approximate moments to approximate the true solution. EM can be carried out by setting $\gamma_i^{t+1} = \langle x_i^2 \rangle$ and iterating over t. After convergence of γ_i's, the mean point estimate is obtained as $\hat{\boldsymbol{x}}_{mean} = \langle x_i \rangle$. The mode point estimate $\hat{\boldsymbol{x}}_{mode}$ can be calculated by using converged values of γ_i's in Eq. (30).

If the diagonal elements of $\boldsymbol{\Sigma}$ are large valued or become large over EM iterations as compared to off-diagonals, then DA is expected to work well. Note that assuming a diagonal

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\boldsymbol{\Sigma}$ was also motivated by previous work [71], [74], [75], [84], [85] for various applications. In this work, we empirically report that DA has very good sparse recovery performance and has low complexity.

To further support the DA approximation, we present empirical findings regarding the structure of $\boldsymbol{\Sigma}$. We performed sparse recovery simulations using Eq. (1) with the MCMC-EM approach as the ground truth (i.e. without regularizing the MCMC estimates of $\hat{\boldsymbol{\Sigma}}$). We assumed that $\boldsymbol{x}$ was of size 200 with 10 non-zero elements drawn from $\mathcal{N}^R(0, 1)$. The dictionary $\boldsymbol{\Phi} \in \mathbb{R}^{50\times 200}$ columns were normally distributed $\boldsymbol{\Phi} \sim \mathcal{N}(0, \boldsymbol{I})$. We solved this problem for 1,000 simulations and overlay plots of the average absolute value of the off-diagonals of $\hat{\boldsymbol{\Sigma}}$ as a function of MCMC-EM iteration in the first row of Fig. 1.

We see that the average of off-diagonal elements decreased exponentially as a function of MCMC-EM iteration, and the average of this behavior over 1,000 simulations (e.g. red line) has a final value of 10^{-4} after 10 iterations. This indicates the off-diagonals of $\hat{\boldsymbol{\Sigma}}$ of the true posterior (with MCMC sampling) indeed become close to zero. Moreover, in the second row of Fig. 1, we overlay plots of the Frobenius norm of the difference between $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}_D$, where $\hat{\boldsymbol{\Sigma}}_D$ is the diagonal matrix consisting of diagonal elements from $\hat{\boldsymbol{\Sigma}}$. This shows that as MCMC-EM converges, the true $\boldsymbol{\Sigma}$ becomes close to a diagonal matrix. These results suggest that, if there is flexibility in choosing the dictionary $\boldsymbol{\Phi}$ as in compressed sensing, then proper choice of $\boldsymbol{\Phi}$ can lead to the DA approach producing high quality approximate marginals $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma})$ that are close to the true marginals.

C. Computational complexity of proposed methods

For computational comparisons, we assume that $N \leq M$. Under this assumption, the time complexity of the DA algorithm is $\mathcal{O}(N^2 M)$ per EM iteration. This complexity is similar to the original SBL algorithm in [44], [52] and is largely due to the computationally intensive matrix inversion step e.g. $(\sigma^2 \boldsymbol{I} + \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^T)^{-1}$ given in Eq. (23). Time complexity of the LMMSE algorithm is also $\mathcal{O}(N^2 M)$ per EM iteration. This complexity is determined from a similar matrix inversion step $(\boldsymbol{\Phi}\boldsymbol{R}_{\boldsymbol{x}}\boldsymbol{\Phi}^T + \sigma^2 \boldsymbol{I})^{-1}$ in Eq. (32) (e.g. note that $\boldsymbol{R}_{\boldsymbol{x}}$ is diagonal). The GAMP algorithm bypasses the computationally intensive matrix inversion and the resulting complexity is $\mathcal{O}(NM)$ time [51]. This is linear in both problem dimensions and significantly faster than the both the DA and LMMSE methods. For the MCMC-EM algorithm, the actual computational cost is determined by the random Hamiltonian MCMC sampling, which is explained in more detail in [65].

IV. Experiment Design

In this section, we provide the layout of our numerical experiments. In the following, we provide extensive comparisons between the proposed R-SBL variants LMMSE, GAMP, MCMC and DA to the baseline S-NNLS solvers, including NNGM-AMP [58], SLEP-ℓ_1 [86], and NN-OMP [87]. In all of the experiments below, we generate sparse vectors $\boldsymbol{x}^{\boldsymbol{gen}} \in \mathbb{R}_+^{400}$, such that $\|\boldsymbol{x}^{\boldsymbol{gen}}\|_0 = K$ and random dictionaries $\boldsymbol{\Phi} \in \mathbb{R}^{100\times 400}$. We normalize the columns of $\boldsymbol{\Phi}$ by $1/\sqrt{N}$ as in [88]. For a fixed $\boldsymbol{\Phi}$ and $\boldsymbol{x}^{gen}$, we compute the measurements $\boldsymbol{y}$ =

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$\boldsymbol{\Phi}\, \boldsymbol{x}^{gen}$ and use the baseline algorithms and the proposed R-SBL variants to approximate $\boldsymbol{x}^{gen}$.

In the first set of experiments, we simulate a 'noiseless' recovery scenario, where the noise variance is set as $\sigma^2 = 10^{-6}$, the non-zero entries of the solution vector are drawn from a rectified Gaussian density $\mathcal{N}^R(0, 1)$ and the dictionary columns are i.i.d. Normal distributed as $\boldsymbol{\Phi} \sim \mathcal{N}(0, \boldsymbol{I})$. We experiment with cardinalities $K = \{10, 20, 30, 35, 40, 45, 50\}$.

In the second set of experiments, we construct various dictionary types to analyze the robustness of our R-SBL method and the baseline solvers for the general S-NNLS problem. The dictionary structures considered here are similar to the ones used in [51], [89] are not necessarily i.i.d. Gaussian, and can be low-rank, coherent, ill-posed, and non-negative as described below:

A. *Coherent dictionaries:* We introduce coherence among the columns of an original dictionary $\boldsymbol{\Phi} = \mathcal{N}(0, \boldsymbol{I})$ and report recovery performances for a fixed $K = 50$. This is done by multiplying $\boldsymbol{\Phi}$ with a coherence matrix $\boldsymbol{C}$, to obtain a dictionary $\boldsymbol{\Phi}_C$ with coherent columns. $\boldsymbol{C}$ is the Cholesky factor of the Toeplitz(ρ) matrix with a coherence parameter ρ. We experimented with different coherence values by selecting $\rho = \{0.1, 0.2, \ldots, 0.80, 0.85, 0.90, 0.95\}$.

B. *Low-rank dictionaries:* We construct rank-deficient dictionaries such that, $\boldsymbol{\Phi} = \boldsymbol{AB}$, where $\boldsymbol{A} \in \mathbb{R}^{N \times R}$, $\boldsymbol{B} \in \mathbb{R}^{R \times M}$ and $R < N$. The elements of $\boldsymbol{A}$ and $\boldsymbol{B}$ are i.i.d. Normal. The rank ratio R/N is considered as a measure of rank deficiency, where smaller values indicate more deviation from an i.i.d. dictionary. We experiment with $R/N = \{1, 0.95, \ldots, 0.4\}$ and report recovery performances for a fixed $K = 50$.

C. *Ill-conditioned dictionaries:* We experiment with ill-conditioned dictionaries with a condition number $\kappa > 1$. For a fixed κ, the dictionary is constructed as $\boldsymbol{\Phi} = \boldsymbol{USV}^T$. Here, $\boldsymbol{U}$ and $\boldsymbol{V}$ contain the left and right singular vectors of an i.i.d. Gaussian matrix, and $\boldsymbol{S}$ is a diagonal matrix containing the eigenvalues. We decay the elements of $\boldsymbol{S}$ with $\boldsymbol{S}_{i+1,i+1} = \kappa^{-1/N-1}\, \boldsymbol{S}_{i,i}$ for $i = 1, 2, \ldots, N-1$. The value of κ measures the deviation from an i.i.d. Gaussian dictionary. We experimented for the condition numbers $\kappa = \{8, 10, \ldots, 28\}$. Larger κ values indicate more deviation from an i.i.d. dictionary.

D. *Non-negative dictionaries:* Non-negative dictionaries are used in sparse recovery applications such as sparse NMF [3] and NN K-SVD [20] where a positive mapping is required on the solution vector. We construct non-negative dictionaries $\boldsymbol{\Phi}$, with columns that are drawn according to $\boldsymbol{\Phi} \sim RG(0, \boldsymbol{I})$. We experiment with cardinalities $K = \{10, 20, 30, 35, 40, 45, 50\}$.

In the third set of experiments, we set the noise variance σ^2 for $\boldsymbol{v}$ such that the signal to noise ratio (SNR) is 20 dB and repeat the first set of experiments. This experiment was meant to assess the robustness of R-SBL variants under noisy conditions.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

In the fourth set of experiments, we investigate recovery performance for a variety of different distributions for non-zero elements of $\boldsymbol{x}$ and different $\boldsymbol{\Phi}$. We draw the nonzero elements of $\boldsymbol{x}^{\boldsymbol{gen}}$ randomly, according to the following probability distributions:

I. NN-Cauchy (Location: 0, Scale: 1)

II. NN-Laplace (Location: 0, Scale: 1)

III. Gamma (Location: 1, Scale: 2)

IV. Chi-square with $\nu = 2$

V. Bernoulli with $p(0.25) = 1/2$ and $p(1.25) = 1/2$

where the prefix 'NN' stands for non-negative. These distributions are obtained by taking the absolute value of the respective probability densities. We generate random dictionaries according to the following densities:

I. Normal (Location: 0, Scale: 1)

II. ±1 with $p(1) = 1/2$ and $p(-1) = 1/2$

III. {0, 1} with $p(0) = 1/2$ and $p(1) = 1/2$

In all of the experiments detailed here, the results we present were averaged over 1,000 simulations. Moreover, the R-SBL MCMC approach was only used in the first set of experiments to demonstrate the high quality of the parameter estimates obtained with the low complexity approaches such as DA, LMMSE and GAMP. We omit the MCMC in other experiments due to computational constraints.

A. Performance metrics

To evaluate the performance of various S-NNLS algorithms, we used the normalized mean square recovery error (NMSE) and the probability of error in the recovered support set (PE) [22]. We computed the NMSE between the recovered signal $\hat{\boldsymbol{x}}$ and the ground truth $\boldsymbol{x}^{gen}$ using:

$$\text{NMSE} = \|\hat{\boldsymbol{x}} - \boldsymbol{x}^{\boldsymbol{gen}}\|^2 / \|\boldsymbol{x}^{\boldsymbol{gen}}\|^2 . \tag{49}$$

The PE metric was computed using:

$$\text{PE} = \frac{max\{|S|, |\hat{S}|\} - |S \cap \hat{S}|}{max\{|S|, |\hat{S}|\}}, \tag{50}$$

where the support of the true solution was S and the support of $\hat{\boldsymbol{x}}$ was $\hat{S}$. A value of PE = 0 indicates that the ground truth and recovered supports are the same, whereas PE = 1 indicates no overlap between supports. Averaging the PE over multiple trials gives the empirical probability of making errors in the recovered support. The averaged values of NMSE and PE over 1,000 simulations for each experiment are reported in the experiment results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. MCMC implementation

We used the MCMC implementation presented in [65]. The Matlab and R codes are available at https://github.com/aripakman/hmc-tmg. MCMC parameters explained in Section III-B1 were selected as follows. The off-diagonal pruning of the empirical scale parameter $\hat{\boldsymbol{\Sigma}}$ was performed with a threshold of $T_p = 5 \times 10^{-2}$. Diagonal scaling was performed with a factor of $\beta = 1.7$, and a shrinkage parameter of $\lambda = 0.5$. These values were empirically determined to minimize the NMSE for the first set of experiments.

V. Experiment Results

Here, we show that in all of the sparse recovery experiments detailed above, the proposed R-SBL variants perform better than the baseline solvers in terms of NMSE and PE. The R-SBL variants outperform the baseline solvers when the dictionary is non i.i.d., coherent, low-rank, ill-posed or even non-negative, illustrating the robustness of R-SBL to different characteristics of the dictionary $\boldsymbol{\Phi}$.

In Fig. 2 (a), we show the sparse recovery performance of the R-SBL variants and the baseline solvers as a function of the cardinality for the first set of experiments. As the cardinality of the ground truth solution increases (e.g. after $K = 30$) the performances of NN-OMP and SLEP-ℓ_1 deteriorate both in terms of MSE and PE. On the other hand, R-SBL variants and NNGM-AMP are quite robust with very small recovery error. At the largest cardinality of $K = 50$, we see that R-SBL DA and MCMC outperform other methods. The DA variant is nearly identical to MCMC in terms of MSE and PE. This is expected since MCMC prunes off-diagonal elements of the scale matrix $\boldsymbol{\Sigma}$ iteratively, when they drop below a certain threshold.

A. Coherent Dictionaries

In Figure 2 (b), we show the recovery performance of the proposed R-SBL variants and the baseline solvers when the dictionary is coherent. The degree of dictionary coherence is shown on the horizontal axis with ρ, which ranges from 0.1 to 0.95. The proposed R-SBL variants are extremely robust to increasing coherence and outperform the baseline solvers in terms of both NMSE and PE. SLEP-ℓ_1 is robust to increasing coherence, but performs worse as compared to R-SBL variants. NNGM-AMP breaks down after $\rho = 0.3$ and performs worse than SLEP-ℓ_1 after $\rho = 0.5$ and than NN-OMP after $\rho = 0.8$. The LMMSE and DA variants are not affected by the coherence level and achieve better recovery even for $\rho = 0.95$. The performance of R-SBL GAMP slightly deteriorates after an extreme coherence of $\rho = 0.90$, but is still better than the baseline solvers.

These results demonstrate that the proposed R-SBL variants are robust to dictionary coherence and are superior to the baseline solvers. The robustness our R-SBL framework seems to be inherited from the robustness of the original SBL algorithm to the structure of $\boldsymbol{\Phi}$ [51], [90], which uses a GSM prior on $\boldsymbol{x}$. Our R-GSM prior on $\boldsymbol{x}$ seems to provide a similar robustness to the R-SBL algorithm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

B. Low-rank Dictionaries

In Figure 2 (c), we show the recovery performance of the proposed R-SBL variants and the baseline solvers for rank-deficient dictionaries. The degree of rank deficiency is shown on the horizontal axis with the rank ratio R/N. The R-SBL variants outperform the baseline solvers both in terms of NMSE and PE for all values of R/N. The recovery performances of the R-SBL variants are extremely robust against the changes in R/N. Among the R-SBL variants, DA performs slightly better than LMMSE and GAMP. The GAMP performs similar to LMMSE. The recovery performance of NNGM-AMP is better than NN-OMP and SLEP-ℓ_1, however its performance degrades as R/N gets smaller.

C. Ill-conditioned Dictionaries

In Figure 3 (a), we demonstrate the recovery performance for ill-conditioned dictionaries. The condition number on the horizontal axis varies from $\kappa = 8$ to $\kappa = 28$. The proposed R-SBL variants perform significantly better than the baseline solvers across different κ values in terms of NMSE and PE. The recovery performance of the R-SBL variants is also extremely robust to different selections of κ. The SLEP-ℓ_1 is better than NN-OMP and NNGM-AMP in terms of recovery performance, and is also robust to the selection of κ. The NN-OMP and NNGM-AMP approaches rapidly deteriorate in recovery performance with increasing κ values.

D. Non-negative Dictionaries

In certain sparse recovery applications, such as sparse NMF and non-negative K-SVD, a positive mapping is desired on the solution vector and the dictionary may contain non-negative elements. In Figure 3 (b), we demonstrate the recovery performance of the S-NNLS solvers when the dictionary is non-negative with elements drawn from i.i.d. $RG(0, 1)$. The cardinality K on the horizontal axis of Figure 3 (b) varies from $K = 10$ to $K = 50$. The NNGM-AMP approach was not able to recover a solution for non-negative dictionaries and point estimates for $\boldsymbol{x}$ diverged for different K. Therefore, the NMSE values for NNGM-AMP were not shown in Figure 3 (b). Unlike in Figure 2 (a), where the dictionary can be both positive and negative, NN-OMP performs better than the SLEP-ℓ_1 solver. The proposed R-SBL variants outperform the baseline approaches. Among the R-SBL variants, DA performs slightly better than the GAMP, and GAMP is slightly better than LMMSE.

E. Noisy Conditions

We compared the recovery performance of the proposed R-SBL variants and the baseline solvers in a noisy setting, where the dictionary is i.i.d. Normal distributed. In this case, the observations were contaminated with additive white Gaussian noise to have a signal to noise ratio (SNR) of 20 dB. Figure 3 (c) shows the NMSE and PE versus the cardinality. Compared with the noiseless case in Figure 2 (a), the performances of all of the methods noticeably reduced. However, the proposed R-SBL variants performed better as compared to the NN-OMP and SLEP-ℓ_1 solvers, and performed similar to the NNGM-AMP approach.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

F. Other types of x^{gen} and Φ

In this fourth set of experiments, the dictionary $\mathbf{\Phi}$ was drawn according to i.i.d. Normal, ±1 Bernoulli, and {0, 1} Bernoulli distributions. We experimented with different distributions for the non-zero elements of $\boldsymbol{x}^{gen}$, as shown in Tables II, III and IV.

For i.i.d. Normal $\mathbf{\Phi}$ in Table II, the R-SBL DA generally outperforms the baseline solvers and other R-SBL variants when $\boldsymbol{x}^{gen}$ is RG, NN-Cauchy, NN-Laplace, Gamma and Chi-square distributed. The LMMSE variant achieves slightly better performance in terms of PE for the NN-Cauchy distribution. The NNGM-AMP is better than LMMSE and GAMP variants, when $\boldsymbol{x}^{gen}$ is RG, however it fails in terms of PE when $\boldsymbol{x}^{gen}$ is NN-Cauchy. The NNGM-AMP approach shows better performance when $\boldsymbol{x}^{gen}$ is Bernoulli. This is expected since the prior density for NNGM-AMP is a Bernoulli non-negative Gaussian mixture. The R-GSM prior, on the other hand, is not well matched to the Bernoulli distribution, as it is a mixture of continuous distributions. Overall, we see that R-SBL DA approach results in the best recovery performance. In Table III, we present the results for when $\mathbf{\Phi}$ is ±1 Bernoulli. The recovery performances observed in Table III are very similar to Table II and overall, R-SBL DA approach enjoys better recovery performance.

In Table IV, we show recovery results for {0, 1} Bernoulli distributed $\mathbf{\Phi}$. The R-SBL DA and LMMSE variants achieve superior recovery when compared to the baseline solvers. The NNGM-AMP approach diverges for different $\boldsymbol{x}^{gen}$, because the dictionary was non-negative, which is consistent with our previous observation when the dictionary elements were drawn from i.i.d. $RG(0, 1)$ in Figure 3 (b).

G. Recovery time analysis

In Section III-C, we presented the worst case computational complexity of the DA, LMMSE and GAMP variants per EM iteration. As the execution time also depends on how fast an EM approach converges to the final solution, we provide an analysis of average execution times for different cardinality values. First, we provide a simple way to speed up the proposed R-SBL algorithms. We prune the problem size, when the elements of $\boldsymbol{\gamma}$ become smaller than a given threshold. For example, when an index of the vector $\boldsymbol{\gamma}$ becomes smaller than i.e. $\gamma_i \leq \varepsilon_{\boldsymbol{\gamma}}$, we ignore the computations regarding that index in the next iterations. This effectively reduces the problem dimensions and improves execution time.

In Fig. 4, we included the average execution times of the proposed algorithms, in units of seconds. The pruning threshold was selected as $\varepsilon_{\boldsymbol{\gamma}} = 10^{-5}$ for all methods. For the EM based methods, we monitored the convergence of the $\boldsymbol{\gamma}$'s in EM iterations. We stopped the EM updates, when $\|\boldsymbol{\gamma}^t - \boldsymbol{\gamma}^{t-1}\|_2 \leq 10^{-3}$, where t is the current EM iteration index. For other approaches, we monitored the linear equality constraints and stopped the algorithms when $\|\boldsymbol{y} - \mathbf{\Phi}\hat{\boldsymbol{x}}^t\|_2 \leq 10^{-3}$, where $\hat{\boldsymbol{x}}^t$ is the solution estimate at iteration t.

As expected due to computationally intensive random sampling, R-SBL MCMC is the slowest method. For display purposes, we scaled down the average MCMC execution time values by 30. The LMMSE approach takes about 3 seconds for $K = 50$ to recover the optimal solution and is the second slowest method. Even though the complexity of DA and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

LMMSE is similar, DA achieves much faster convergence and takes about 0.5 to 1 seconds as K increases.

For this particular experiment, GAMP is the fastest R-SBL variant regardless of the cardinality and is similar to SLEP-ℓ_1. However, since complexity of GAMP is $\mathcal{O}(NM)$, for very large problem sizes (e.g. large N and M) GAMP may become slower despite the superior recovery performance. In this case, a convex solver may be preferable depending on the desired recovery performance. R-SBL GAMP is faster than NNGM-AMP at larger cardinalities. Finally, NN-OMP is similar to SLEP-ℓ_1 but it's execution time increases for larger cardinalities. Considering the fast recovery speed and good recovery performance of R-SBL GAMP under various $\mathbf{\Phi}$ types, GAMP is a very good candidate for time sensitive sparse recovery applications.

H. Application on real data: Face Recognition

Here, we present a face recognition (FR) application, based on the non-negative sparse representations considered in [91], [92], [93]. Our goal is to show that the R-SBL approach works in real-world applications involving real-data. A sparse representation classifier (SRC) for FR was initially proposed in [94] using the ℓ_1 penalty without the non-negativity constraints. The SRC approach was found to be robust against occlusion, disguise, pixel corruptions, and achieved superior results as compared to well-known FR algorithms [94], [92], [95], [96].

In the SRC framework, the dictionary $\mathbf{\Phi}$ represents the training samples, and each column of $\mathbf{\Phi}$ contains training features from a single face image. A single person might have more than one training image, and hence multiple columns of $\mathbf{\Phi}$ might correspond to the same person. For a given test face $\boldsymbol{y}$ in vectorized form, a vector $\boldsymbol{x}$ is obtained by solving Eq. (1), using ℓ_1 sparsity, with the assumption that only a few non-zero entries will exist in the solution $\boldsymbol{x}$. Ideally, the index of maximal non-negative entry in $\boldsymbol{x}$ is used to select the corresponding column in $\mathbf{\Phi}$. This should correspond to one of the training samples for the correct person. In [92], the SRC performance was further improved, by adding the non-negativity constraint on $\boldsymbol{x}$ in addition to the ℓ_1 sparsity. The authors have showed their algorithm to be more robust against noise, and to be computationally more effective as compared to the original SRC approach.

In our experiment, we consider the R-SBL framework for the FR problem and compare it with the baseline solvers. Note that, SLEP-ℓ_1 was considered as the non-negative ℓ_1 minimization counterpart of R-SBL in place of [92]. We used the public AR dataset [97], and selected the first 30 males and 30 females for the FR problem. Each person in the dataset has 26 face images with different facial expression, illumination and disguise (e.g. sunglasses and scarves). The first 13 images of each person ($M = 13 \times 60 = 780$) were selected as the training set, and the remaining 780 face images were used for testing. For feature selection, we used the down-sampling method used in [91], [92], [94], where the pixel dimensions of each face image were down-sampled to have a total of N pixels. In separate experiments, each 165×120 pixel image was down-sampled by a factor of {1/28, 1/26, …, 1/6}, yielding feature dimensions of minimum of $N = 30$ to a maximum of $N = 650$.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The overall process is shown in Figure 5 (a), where a query face is shown in the top right-hand side. This image was down-sampled and the original feature dimension was reduced from 19, 800 to 512. After sparse recovery with R-SBL, the original faces belonging to several largest non-zero elements of $\boldsymbol{x}$ are shown. As desired, the maximal positive index of $\boldsymbol{x}$ belongs to the same person in the query face.

In Figure 5 (b), we performed FR using all 780 samples in the test set and measured the recognition rate for different feature sizes. The recognition rate was computed by counting the number of test samples for which R-SBL recovered the correct individual from $\boldsymbol{x}$. This count was normalized by 780. Overall, the R-SBL variants with the exception of R-SBL GAMP performed similar to the baseline solvers for large feature sizes. This is expected since the recovery problem was highly sparse, and the cardinality was very small $K = 13$ as compared to the length of $\boldsymbol{x}$ e.g. largest length of $\boldsymbol{x}$ is 780. R-SBL GAMP was superior to all algorithms for large feature sizes and performed significantly better in identifying the correct individual. NNGM-AMP diverged for this application and did not yield reportable results.

VI. Conclusion

In this work, we introduced a hierarchical Bayesian method to solve the S-NNLS problem. We proposed the rectified Gaussian scale mixture model as a general and versatile prior to promote sparsity in the solution of interest. Since the marginals of the posterior were not tractable, we constructed our R-SBL algorithm using the EM framework using four different approaches. We demonstrated that our R-SBL approaches outperformed the available S-NNLS solvers, in most cases by a large margin. The proposed R-SBL framework is very robust to the structure of $\boldsymbol{\Phi}$ and performed well regardless of $\boldsymbol{\Phi}$ being i.i.d. and non-i.i.d. The performance gains achieved by R-SBL variants are consistent across different non-negative data distributions for $\boldsymbol{x}$, different structures for the design matrix $\boldsymbol{\Phi}$, in coherent, low-rank, ill-posed and non-negative settings. The DA variant was found to be a fairly easy to implement S-NNLS solver with simple closed-form moment expressions.

Acknowledgments

We would like to thank Mr. Sung-En Chiu for his comments on an earlier version of this manuscript.

This work was partially supported by NIH grant R21MH112155.

Appendix

A. Full derivation of GAMP

We use the R-GSM prior $p(\boldsymbol{x}|\boldsymbol{\gamma})$ and evaluate Eq. (39) and Eq. (40) to find the first two moments of the approximate marginal posterior under the sum-product GAMP mode

$$\hat{x}_i = \mathbb{E}\{x_i | r_i; \tau_{r_i}\} = \int_{x_i} x_i p(x_i | r_i; \tau_{r_i}) \tag{51}$$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$= \int_{+} x_i \mathcal{N}^R(x_i|0,\gamma_i)\mathcal{N}(x_i, r_i, \tau_{r_i}), \tag{52}$$

then using Gaussian multiplication rule[2], we obtain

$$\hat{x}_i = \int_{+} x_i \Upsilon \mathcal{N}^R(x_i|\eta_i, \nu_i), \tag{53}$$

where η_i and ν_i are given in Eq. (55) and Eq. (56).

We find the mean of the resulting rectified Gaussian

$$\hat{x}_i = \eta_i + \sqrt{\nu_i} h(\frac{\eta_i}{\nu_i}) \tag{54}$$

$$\eta_i = \frac{r_i \gamma_i}{\tau_{r_i} + \gamma_i} \tag{55}$$

$$\nu_i = \frac{\tau_{r_i} \gamma_i}{\tau_{r_i} + \gamma_i} \tag{56}$$

$$h(a) = \frac{\varphi(a)}{\Phi_c(a)}, \tag{57}$$

where φ refers to the pdf and Φ_c refers to the complementary cdf of a zero-mean and unit-variance Gaussian distribution. The conditional variance of x_i given r_i is simply

$$\tau_{x_i} = var\{x_i|r_i;\tau_{r_i}\} = \int_{x_i} x_i^2 p(x_i|r_i;\tau_{r_i}) - \hat{x}_i^2 \tag{58}$$

$$= \int_{+} x_i^2 \mathcal{N}^R(x_i|0,\gamma_i)\mathcal{N}(x_i, r_i, \tau_{r_i}) - \hat{x}_i^2, \tag{59}$$

using Gaussian pdf multiplication rule

$$\tau_{x_i} = \int_{+} x_i^2 \Upsilon \mathcal{N}^R(x_i|\eta_i, \nu_i), \tag{60}$$

we find the variance of the resulting rectified Gaussian

[2] $\mathcal{N}(x;\mu_a,\tau_a)\mathcal{N}(x;\mu_b,\tau_b) = \Upsilon\mathcal{N}(x; \frac{\frac{\mu_a}{\tau_a}+\frac{\mu_b}{\tau_b}}{\frac{1}{\tau_a}+\frac{1}{\tau_b}}, \frac{1}{\frac{1}{\tau_a}+\frac{1}{\tau_b}})$, where Υ is a scaling factor.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$\tau_{x_i} = \nu_i g(\frac{\eta_i}{\nu_i}) \tag{61}$$

$$g(a) = 1 - h(a)(h(a) - a)\,. \tag{62}$$

In the case of max-sum GAMP implementation, we evaluate Eq. (41) and Eq. (42)

$$\hat{x}_i = \arg\min_{\hat{x}_i \geq 0} \frac{x_i^2}{2\gamma_i} + \frac{1}{2\tau_{r_i}} |\hat{x}_i - r_i|^2 \tag{63}$$

$$\hat{x}_i = \begin{cases} \dfrac{r_i \gamma_i}{\tau_{r_i} + \gamma_i} = \eta_i & \text{if } \hat{x}_i \geq 0 \\ 0 & \text{if } \hat{x}_i < 0 \end{cases} \tag{64}$$

Using Eq. (42)

$$\tau_{x_i} = \begin{cases} \dfrac{\tau_{r_i} \gamma_i}{\tau_{r_i} + \gamma_i} = \nu_i & \text{if } \hat{x}_i \geq 0 \\ 0 & \text{if } \hat{x}_i < 0^3 \end{cases} \tag{65}$$

Upon convergence of the max-sum, the approximate marginals are obtained using Eq. (54) and Eq. (61).

B. Approximate marginals and moments using DA

We derive the approximate moments used in the R-SBL DA approximation. We start with the posterior $p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma})$ and use chain rule to write

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma}) = \frac{p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\gamma}) p(\boldsymbol{x} \mid \boldsymbol{\gamma})}{\int_{\boldsymbol{x}} p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\gamma}) p(\boldsymbol{x} \mid \boldsymbol{\gamma}) d\boldsymbol{x}}\,. \tag{66}$$

Here $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\gamma})$ is a Gaussian density due to the Gaussian noise assumption. Since $p(\boldsymbol{x} \mid \boldsymbol{\gamma})$ is a rectified Gaussian density the numerator of Eq. (66) is a Gaussian multiplied by a rectified Gaussian, which results in a rectified Gaussian density. Then, we can simply write

$$p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma}) = c(\boldsymbol{y}) e^{-\frac{(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})}{2}} u(\boldsymbol{x}), \tag{67}$$

where $c(\boldsymbol{y})$ is the normalizing constant for the posterior density and $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are given by Eqs. (23) and (24), respectively. Let $\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$ and $\boldsymbol{r} = \boldsymbol{x} - \boldsymbol{\mu}$, so that $d\boldsymbol{x} = d\boldsymbol{r}$ and $\boldsymbol{\Sigma}^{-1} = \boldsymbol{L}^{-T}\boldsymbol{L}^{-1}$. Therefore, we have

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$1 = c(\boldsymbol{y}) \int_{-\boldsymbol{\mu}}^{\infty} e^{\frac{-\boldsymbol{r}^T \boldsymbol{L}^{-T} \boldsymbol{L}^{-1} \boldsymbol{r}}{2}} d\boldsymbol{r}\,. \tag{68}$$

Now, let $\boldsymbol{z} = \boldsymbol{L}^{-1}\boldsymbol{r}$, which implies that $d\boldsymbol{r} = |\boldsymbol{L}| d\boldsymbol{z}$ and

$$c(\boldsymbol{y}) = \frac{1}{|\boldsymbol{L}| \int_{-\boldsymbol{\beta}}^{\infty} e^{-\boldsymbol{z}^T \boldsymbol{z}/2} d\boldsymbol{z}}, \tag{69}$$

where $\boldsymbol{\beta} = \boldsymbol{L}^{-1}\boldsymbol{\mu}$ is the lower limit of the new integral in vector form. The lower limit $\boldsymbol{\beta}$ depends on a linear combination of elements of $\boldsymbol{\mu}$ since $\boldsymbol{L}$ is not diagonal. Thus, the integral in the denominator of Eq. (69) is **not tractable** as the integration limits are not separable and the multidimensional integral over $\boldsymbol{z}$ in Eq. (69) is not separable as a product of one dimensional integrals.

Assume that, we are interested in an approximate density $\tilde{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma})$, instead of the exact posterior. We calculate an approximate $\tilde{c}(\boldsymbol{y})$ by approximating $\boldsymbol{\Sigma}$ with its diagonal i.e. $\boldsymbol{\Sigma}_d = diag(\boldsymbol{\Sigma}) \approx \boldsymbol{\Sigma}$. In this case, the new $\boldsymbol{L}$ is diagonal with entries $\sqrt{\sum_{ii}}$. Thus, the integral in Eq. (69) is separable and the approximate normalizing constant $\tilde{c}(\boldsymbol{y})$ has closed form

$$\tilde{c}(\boldsymbol{y}) = \frac{1}{|\boldsymbol{\Sigma}_d|^{1/2} \prod_{i=1}^{M} \sqrt{\frac{\pi}{2}} \operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)}. \tag{70}$$

Approximating the actual normalizing constant with $\tilde{c}(\boldsymbol{y})$, we write the approximate posterior as

$$\tilde{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma}) = \frac{e^{-\frac{(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}_d^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}{2}} u(\boldsymbol{x})}{\prod_{i=1}^{M} \sqrt{\frac{\pi \sum_{ii}}{2}} \operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)} \tag{71}$$

$$= \prod_{i=1}^{M} \sqrt{\frac{2}{\pi \sum_{ii}}} \frac{e^{-\frac{(x_i-\mu_i)^2}{2\sum_{ii}}} u(x_i)}{\operatorname{erfc}\left(-\frac{\mu_i}{\sqrt{2\sum_{ii}}}\right)} \tag{72}$$

$$= \prod_{i=1}^{M} \tilde{p}(x_i \mid \boldsymbol{y}, \boldsymbol{\gamma}) \tag{73}$$

Eq. (73) shows that multivariate $\tilde{p}(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\gamma})$ is separable into product of univariate densities. The univariate density $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma})$ is the univariate RG density defined in Eq. (10) e.g. $\tilde{p}(x_i|\boldsymbol{y}, \boldsymbol{\gamma}) = \mathcal{N}^R(x_i; \mu_i, \Sigma_{ii})$. The first and second moments of a univariate RG density are well-known in closed form (i.e. Eqs. (47) and (48)) and are used in the R-SBL DA algorithm.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

References

1. Lawson, CL, Hanson, RJ. Solving least squares problems. Vol. 161. SIAM; 1974.
2. Jedynak BM, Khudanpur S. Maximum likelihood set for estimating a probability mass function. Neural Computation. 17(7):1508–1530.2005; [PubMed: 15901406]
3. Peharz R, Pernkopf F. Sparse nonnegative matrix factorization with ℓ_0-constraints. Neurocomputing. 80:38–46.2012; [PubMed: 22505792]
4. Kim H, Park H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics. 23(12):1495–1502.2007; [PubMed: 17483501]
5. Kim H, Park H. Nonnegative matrix factorization based on alternating nonnegativity constrained least squares and active set method. SIAM Journal on Matrix Analysis and Applications. 30(2):713–730.2008;
6. Fedorov I, Nalci A, Giri R, Rao BD, Nguyen TQ, Garudadri H. A unified framework for sparse non-negative least squares using multiplicative updates and the non-negative matrix factorization problem. Signal Processing. 2018
7. Pauca VP, Shahnaz F, Berry MW, Plemmons RJ. Text mining rsing non-negative matrix factorizations. Proceedings of the 2004 SIAM International Conference on Data Mining. 42004; :452–456.
8. Monga V, Mihçak MK. Robust and secure image hashing via non-negative matrix factorizations. IEEE Transactions on Information Forensics and Security. 2(3):376–390.2007;
9. Loizou PC. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Transactions on Speech and Audio Processing. 13(5):857–869.2005;
10. Févotte C, Bertin N, Durrieu J-L. Nonnegative matrix factorization with the Itakura-Saito divergence: With application to music analysis. Neural Computation. 21(3):793–830.2009; [PubMed: 18785855]
11. Sajda P, Du S, Brown TR, Stoyanova R, Shungu DC, Mao X, Parra LC. Nonnegative matrix factorization for rapid recovery of constituent spectra in magnetic resonance chemical shift imaging of the brain. IEEE Transactions on Medical Imaging. 23(12):1453–1465.2004; [PubMed: 15575404]
12. Lin Y, Lee DD. Bayesian regularization and nonnegative deconvolution for room impulse response estimation. IEEE Transactions on Signal Processing. 54(3):839–847.2006;
13. Potter LC, Ertin E, Parker JT, Cetin M. Sparsity and compressed sensing in radar imaging. Proceedings of the IEEE. 98(6):1006–1020.2010;
14. Hurmalainen A, Saeidi R, Virtanen T. Group sparsity for speaker identity discrimination in factorisation-based speech recognition. Interspeech. 2012
15. Lustig M, Santos JM, Donoho DL, Pauly JM. kt SPARSE: High frame rate dynamic MRI exploiting spatio-temporal sparsity. Proceedings of the 13th Annual Meeting of ISMRM. 24202006;
16. Ghosh, A, Megherbi, T, Boumghar, FO, Deriche, R. 10th International Symposium on Biomedical Imaging (ISBI). IEEE; 2013. Fiber orientation distribution from non-negative sparse recovery; 254–257.
17. Meng J, Zhang JM, Chen Y, Huang Y. Bayesian non-negative factor analysis for reconstructing transcription factor mediated regulatory networks. Proteome Science. 9(1):S9.2011; [PubMed: 22166063]
18. Nalci A, Rao B, Liu TT. Sparse Estimation of Quasi-periodic Spatiotemporal Components in Resting-State fMRI. Proceedings of the 24th Annual Meeting of the ISMRM. 2016:824.
19. Liu TT, Nalci A, Falahpour M. The global signal in fmri: Nuisance or information? NeuroImage. 150:213–229.2017; [PubMed: 28213118]
20. Aharon, M, Elad, M, Bruckstein, AM. Optics & Photonics 2005. International Society for Optics and Photonics; 2005. K-SVD and its nonnegative variant for dictionary design; 591 411–591 411.
21. Jiang X, Ye Y. A note on complexity of lp minimization. Preprint. 2009
22. Elad, M. Sparse and Redundant Representations. Springer; New York: 2010.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

23. Tropp JA, Gilbert AC. Signal recovery from random measurements via orthogonal matching pursuit. IEEE Transactions on information theory. 53(12):4655–4666.2007;
24. Needell D, Tropp JA. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. Applied and Computational Harmonic Analysis. 26(3):301–321.2009;
25. Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Transactions on signal processing. 41(12):3397–3415.1993;
26. Needell D, Vershynin R. Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Foundations of computational mathematics. 9(3):317–334.2009;
27. Bruckstein AM, Donoho DL, Elad M. From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Review. 51(1):34–81.2009;
28. Pati, YC, Rezaiifar, R, Krishnaprasad, P. Asilomar Conference on Signals, Systems and Computers. IEEE; 1993. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition; 40–44.
29. Efron B, Hastie T, Johnstone I, Tibshirani R, et al. Least angle regression. The Annals of statistics. 32(2):407–499.2004;
30. Figueiredo MA, Nowak RD, Wright SJ. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. IEEE Journal of selected topics in signal processing. 1(4):586–597.2007;
31. Wright SJ, Nowak RD, Figueiredo MA. Sparse reconstruction by separable approximation. IEEE Transactions on Signal Processing. 57(7):2479–2493.2009;
32. Donoho DL, Tanner J. Sparse nonnegative solution of underdetermined linear equations by linear programming. Proceedings of the National Academy of Sciences of the United States of America. 102(27):9446–9451.2005; [PubMed: 15976026]
33. Nocedal, J, Wright, S. Numerical Optimization. Springer Science & Business Media; 2006.
34. Boyd, S, Vandenberghe, L. Convex Optimization. Cambridge University Press; 2004.
35. Khajehnejad MA, Dimakis AG, Xu W, Hassibi B. Sparse recovery of nonnegative signals with minimal expansion. IEEE Transactions on Signal Processing. 59(1):196–208.2011;
36. Lin, C-b. Projected gradient methods for nonnegative matrix factorization. Neural Computation. 19(10):2756–2779.2007; [PubMed: 17716011]
37. Grady, PD, Rickard, ST. IEEE Workshop on Machine Learning for Signal Processing. IEEE; 2008. Compressive sampling of non-negative signals; 133–138.
38. Chartrand, R, Yin, W. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2008. Iteratively reweighted algorithms for compressive sensing; 3869–3872.
39. Giri R, Rao BD. Type I and Type II Bayesian Methods for Sparse Signal Recovery using Scale Mixtures. IEEE Transactions on Signal Processing. 642016;
40. Tipping ME. Sparse Bayesian learning and the relevance vector machine. The Journal of Machine Learning Research. 1:211–244.2001;
41. Babacan SD, Molina R, Katsaggelos AK. Bayesian compressive sensing using Laplace priors. IEEE Transactions on Image Processing. 19(1):53–63.2010; [PubMed: 19775966]
42. Ji S, Xue Y, Carin L. Bayesian compressive sensing. IEEE Transactions on Signal Processing. 56(6):2346–2356.2008;
43. Andrews DF, Mallows CL. Scale mixtures of normal distributions. Journal of the Royal Statistical Society. Series B (Methodological). :99–102.1974
44. Wipf DP, Rao BD. An empirical Bayesian strategy for solving the simultaneous sparse approximation problem. IEEE Transactions on Signal Processing. 55(7):3704–3716.2007;
45. Palmer, JA. Ph.D. dissertation. University of California; San Diego: 2006. Variational and scale mixture representations of non- Gaussian densities for estimation in the Bayesian linear model: Sparse coding, independent component analysis, and minimum entropy segmentation.
46. Palmer J, Kreutz-Delgado K, Rao BD, Wipf DP. Variational EM algorithms for non-Gaussian latent variable models. Advances in Neural Information Processing Systems. 2005:1059–1066.
47. Lange K, Sinsheimer JS. Normal/independent distributions and their applications in robust regression. Journal of Computational and Graphical Statistics. 2(2):175–198.1993;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

48. Dempster AP, Laird NM, Rubin DB. Iteratively reweighted least squares for linear regression when errors are normal/independent distributed. Multivariate Analysis V. :35–57.1980
49. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B. :1–38.1977
50. Wipf DP, Rao BD, Nagarajan S. Latent variable Bayesian models for promoting sparsity. IEEE Transactions on Information Theory. 57(9):6236–6255.2011;
51. Al-Shoukairi M, Schniter P, Rao BD. A GAMP-based low complexity sparse Bayesian learning algorithm. IEEE Transactions on Signal Processing. 66(2):294–308.2018;
52. Wipf DP, Rao BD. Sparse Bayesian learning for basis selection. IEEE Transactions on Signal Processing. 52(8):2153–2164.2004;
53. Zhang, L, Yang, M, Feng, X. IEEE International Conference on Computer Vision (ICCV). IEEE; 2011. Sparse representation or collaborative representation: Which helps face recognition?; 471–478.
54. Fedorov, I, Rao, BD, Nguyen, TQ. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE; 2017. Multimodal sparse Bayesian dictionary learning applied to multimodal data classification; 2237–2241.
55. Harva M, Kabán A. Variational learning for rectified factor analysis. Signal Processing. 87(3):509–527.2007;
56. Miskin, JW. Advances in Independent Component Analysis. Citeseer; 2000. Ensemble learning for independent component analysis.
57. Figueiredo MA. Adaptive sparseness for supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 25(9):1150–1159.2003;
58. Vila JP, Schniter P. An empirical-Bayes approach to recovering linearly constrained non-negative sparse signals. IEEE Transactions on Signal Processing. 62(18):4689–4703.2014;
59. Gauvain J-L, Lee C-H. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. IEEE Transactions on Speech and Audio Processing. 2(2):291–298.1994;
60. Bishop CM. Pattern Recognition. Machine Learning. 2006
61. Zhang Z, Jung T-P, Makeig S, Pi Z, Rao B. Spatiotemporal sparse Bayesian learning with applications to compressed sensing of multichannel physiological signals. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 22(6):1186–1197.2014; [PubMed: 24801887]
62. Horrace WC. Some results on the multivariate truncated normal distribution. Journal of Multivariate Analysis. 94(1):209–221.2005;
63. Robert, C, Casella, G. Monte Carlo statistical methods. Springer Science & Business Media; 2013.
64. Duane S, Kennedy AD, Pendleton BJ, Roweth D. Hybrid Monte Carlo. Physics Letters B. 195(2):216–222.1987;
65. Pakman A, Paninski L. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. Journal of Computational and Graphical Statistics. 23(2):518–542.2014;
66. Andrieu C, De Freitas N, Doucet A, Jordan MI. An introduction to MCMC for machine learning. Machine Learning. 50(1–2):5–43.2003;
67. Sherman RP, Ho Y-YK, Dalal SR. Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. The Econometrics Journal. 2(2):248–267.1999;
68. Geman S, Geman D. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Transactions on Pattern Analysis and Machine Intelligence. (6):721–741.1984; [PubMed: 22499653]
69. Li Y, Ghosh SK. Efficient sampling methods for truncated multivariate normal and Student-t distributions subject to linear inequality constraints. Journal of Statistical Theory and Practice. 9(4):712–732.2015;
70. Neath, RC, , et al. Advances in Modern Statistical Theory and Applications. Institute of Mathematical Statistics; 2013. On convergence properties of the Monte Carlo EM algorithm; 43–62.
71. Bickel PJ, Levina E. Covariance regularization by thresholding. The Annals of Statistics. :2577–2604.2008

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

72. Tong T, Wang C, Wang Y. Estimation of variances and covariances for high-dimensional data: a selective review. Wiley Interdisciplinary Reviews: Computational Statistics. 6(4):255–264.2014;
73. Celeux G, Diebolt J. A stochastic approximation type EM algorithm for the mixture problem. Stochastics: An International Journal of Probability and Stochastic Processes. 41(1–2):119–134.1992;
74. Fisher TJ, Sun X. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Computational Statistics & Data Analysis. 55(5):1909–1918.2011;
75. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis. 88(2):365–411.2004;
76. Kailath, T, Sayed, AH, Hassibi, B. Linear Estimation. Vol. 1. Prentice Hall Upper Saddle; River, NJ: 2000.
77. Miskin, JW. Advances in Independent Component Analysis. Citeseer; 2000. Ensemble learning for independent component analysis.
78. Rangan, S. International Symposium on Information Theory Proceedings (ISIT). IEEE; 2011. Generalized approximate message passing for estimation with random linear mixing; 2168–2172.
79. Rangan, S, Schniter, P, Fletcher, A. IEEE International Symposium on Information Theory (ISIT). IEEE; 2014. On the convergence of approximate message passing with arbitrary matrices; 236–240.
80. Caltagirone, F, Zdeborová, L, Krzakala, F. IEEE International Symposium on Information Theory (ISIT). IEEE; 2014. On convergence of approximate message passing; 1812–1816.
81. Rangan S, Schniter P, Fletcher A. On the convergence of approximate message passing with arbitrary matrices. IEEE International Symposium on Information Theory (ISIT). Jun.2014 :236–240.
82. Javanmard A, Montanari A. State evolution for general approximate message passing algorithms with applications to spatial coupling. Information and Inference. :iat004.2013
83. Vila JP, Schniter P. Expectation-maximization Gaussian-mixture approximate message passing. IEEE Transactions on Signal Processing. 61(19):4658–4672.2013;
84. Horrace WC. On ranking and selection from independent truncated normal distributions. Journal of Econometrics. 126(2):335–354.2005;
85. Magdon-Ismail, M, Purnell, JT. Intelligent Data Engineering and Automated Learning. Springer; 2010. Approximating the covariance matrix of GMMs with low-rank perturbations; 300–307.
86. Liu J, Ji S, Ye J, et al. SLEP: Sparse learning with efficient projections. Arizona State University. 6:491.2009;
87. Bruckstein AM, Elad M, Zibulevsky M. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. IEEE Transactions on Information Theory. 54(11):4813–4820.2008;
88. Candes EJ, Romberg JK, Tao T. Stable signal recovery from incomplete and inaccurate measurements. Communications on Pure and Applied Mathematics. 59(8):1207–1223.2006;
89. Vila, J, Schniter, P, Rangan, S, Krzakala, F, Zdeborová, L. Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE; 2015. Adaptive damping and mean removal for the generalized approximate message passing algorithm; 2021–2025.
90. Li T, Zhang Z. Robust face recognition via block sparse bayesian learning. Mathematical Problems in Engineering. 20132013;
91. He R, Zheng W-S, Hu B-G, Kong X-W. Two-stage nonnegative sparse representation for large-scale face recognition. IEEE Transactions on Neural Networks and Learning Systems. 24(1):35–46.2013; [PubMed: 24808205]
92. He R, Zheng W-S, Hu B-G. Maximum correntropy criterion for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 33(8):1561–1576.2011; [PubMed: 21135440]
93. Vo N, Moran B, Challa S. Nonnegative-least-square classifier for face recognition. Advances in Neural Networks–ISNN 2009. :449–456.2009

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

94. Wright J, Yang AY, Ganesh A, Sastry SS, Ma Y. Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence. 31(2):210–227.2009; [PubMed: 19110489]

95. Turk M, Pentland A. Eigenfaces for recognition. Journal of Cognitive Neuroscience. 3(1):71–86.1991; [PubMed: 23964806]

96. He X, Yan S, Hu Y, Niyogi P, Zhang H-J. Face recognition using laplacianfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence. 27(3):328–340.2005; [PubMed: 15747789]

97. Martinez AM. The AR face database. CVC technical report. 1998

Biographies

Alican Nalci (S'15) received the B.Sc. degree in Electrical and Electronics Engineering from Bilkent University, Ankara, Turkey, in 2013. He received the M.Sc. degree in Electrical and Computer Engineering from the University of California San Diego, La Jolla, CA, USA, in 2015, where he is currently pursuing a Ph.D. degree. His research interests include sparse signal recovery, machine learning, signal processing and functional magnetic resonance imaging (fMRI).

Igor Fedorov (S'15) receieved the B.Sc. and M.Sc. degrees in Electrical Engineering from the University of Illinois at Urbana-Champaign in 2012 and 2014, respectively. He is currently pursuing a Ph.D. degree in Electrical Engineering at the University of California San Diego, La Jolla, CA, USA. His research interests include sparse signal recovery, machine learning, and signal processing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Maher Al-Shoukairi (S'17) received his B.Sc. degree in Electrical Engineering from the University of Jordan in 2005. He received his M.Sc. degree in Electrical Engineering from Texas A&M University in 2008, after which he joined Qualcomm Inc. on the same year to date. He is currently pursuing his Ph.D. degree in Electrical Engineering at the University of California, San Diego.

Thomas T. Liu received the B.S degree in Electrical Engineering from the Massachusetts Institute of Technology, Cambridge, MA, USA, and the M.S. and Ph.D. degrees from Stanford University, Stanford CA, USA, in 1988, 1993, and 1999, respectively. Since 1999 he has been with the University of California San Diego, La Jolla, CA, USA, where he is currently a Professor in the Departments of Radiology, Psychiatry, and Bioengineering and Director of the UCSD Center for Functional MRI.

Bhaskar D. Rao (S'80-M'83-SM'91-F'00) received the B.Tech. degree in electronics and electrical communication engineering from the Indian Institute of Technology of Kharagpur, Kharagpur, India, in 1979, and the M.S. and Ph.D. degrees from the University of Southern California, Los Angeles, CA, USA, in 1981 and 1983, respectively. Since 1983, he has been at the University of California at San Diego, San Diego, CA, USA, where he is currently a Distinguished Professor in the Department of Electrical and Computer Engineering.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

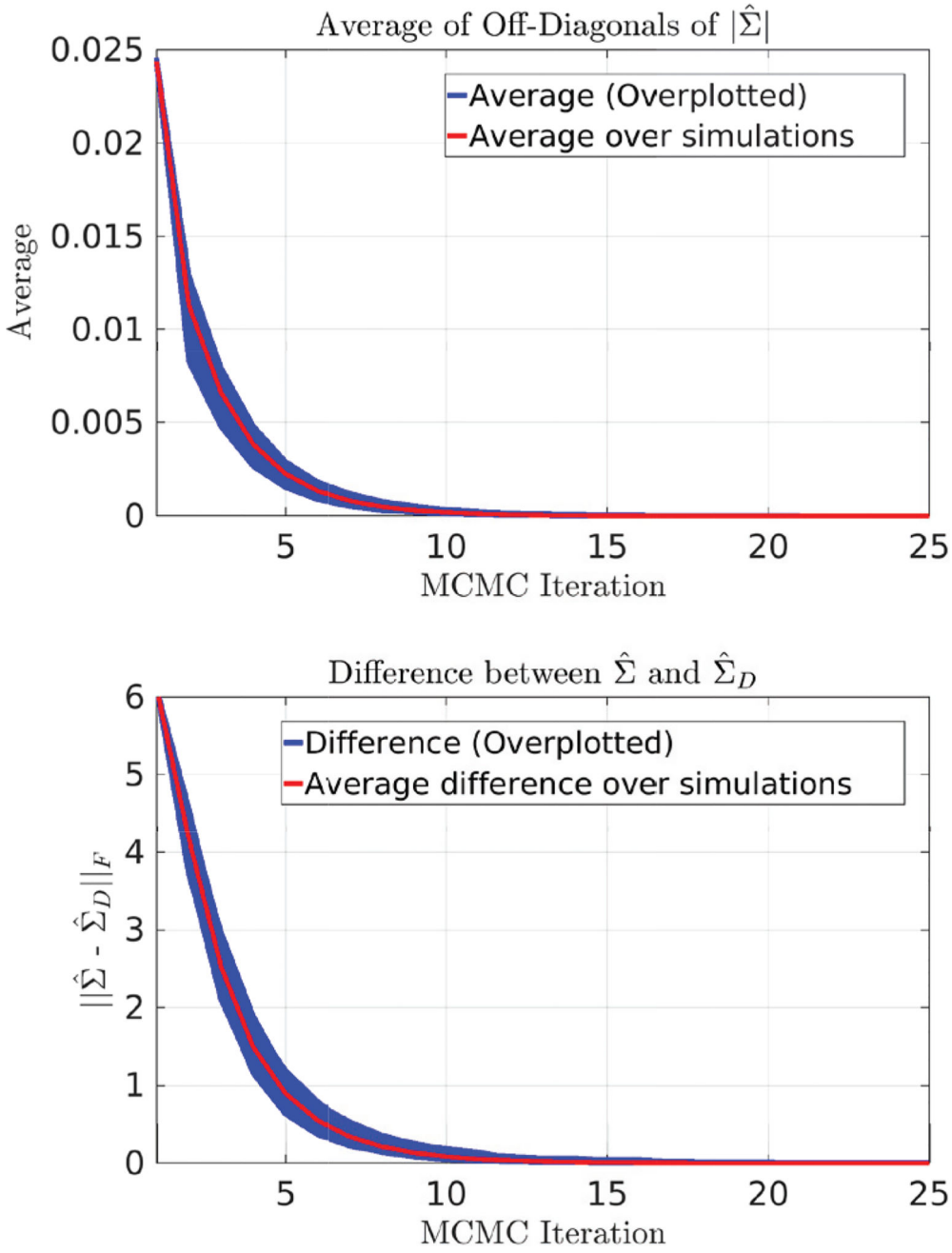

Fig. 1.

Top: Empirical observations for the structure of $\boldsymbol{\Sigma}$. We performed S-NNLS recovery using MCMC-EM (without regularizing the estimates of $\boldsymbol{\Sigma}$) and monitored the average value of off-diagonals for $|\boldsymbol{\Sigma}|$. We simulated for 1,000 runs and overplotted the results (blue lines). The average of average off-diagonals for $|\boldsymbol{\Sigma}|$ over 1,000 results is shown with the red line. The exponentially decreasing behavior suggests that the off-diagonal magnitudes of $\boldsymbol{\Sigma}$ decrease over MCMC iterations, suggesting that true $\boldsymbol{\Sigma}$ is close to a diagonal form. *Bottom*: Distance between true $\boldsymbol{\Sigma}$ and a diagonal matrix formed by it's diagonal entries $\boldsymbol{\Sigma}_D$ for the

same experiment. This suggests that true $\boldsymbol{\Sigma}$ approaches to a diagonal matrix over MCMC iterations.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

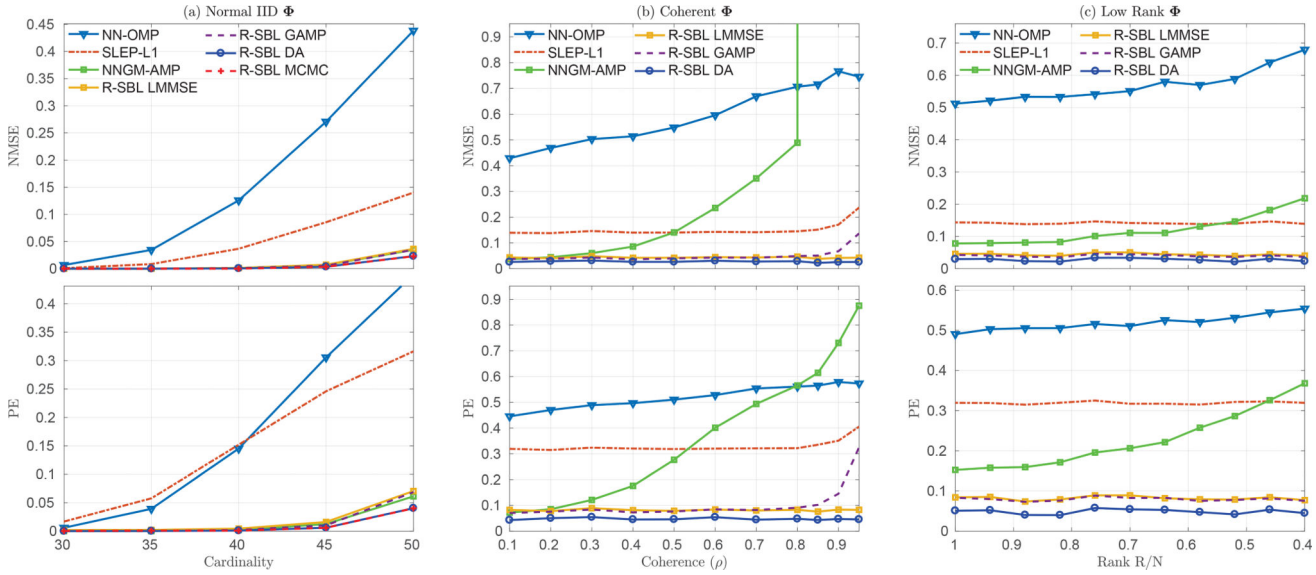

Fig. 2.

Sparse recovery performances (NMSE and PE) of the R-SBL variants and the baseline S-NNLS solvers for various **Φ**. In (a), the dictionary elements were drawn from i.i.d Normal **Φ** and sparse recovery results are shown for cardinalities $K = 30$ to $K = 50$. R-SBL DA achieves the best recovery performance. R-SBL LMMSE and GAMP are similar to NNGM-AMP and are much better than SLEP-ℓ_1 and NN-OMP. In (b), the dictionary columns are coherent with the coherence degree ρ indicated in the x-axis. R-SBL variants are extremely robust to increasing coherence and result in very small NMSE and PE across all ρ values. NNGM-AMP breaks down after $\rho = 0.2$ and its performance deteriorates with increasing ρ. SLEP-ℓ_1 is better than NNGM-AMP after $\rho = 0.5$. In (c), the dictionary is rank-deficient with rank-ratio R/N indicated in the x-axis. R-SBL variants are superior to baseline sparse recovery methods across all R/N values.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

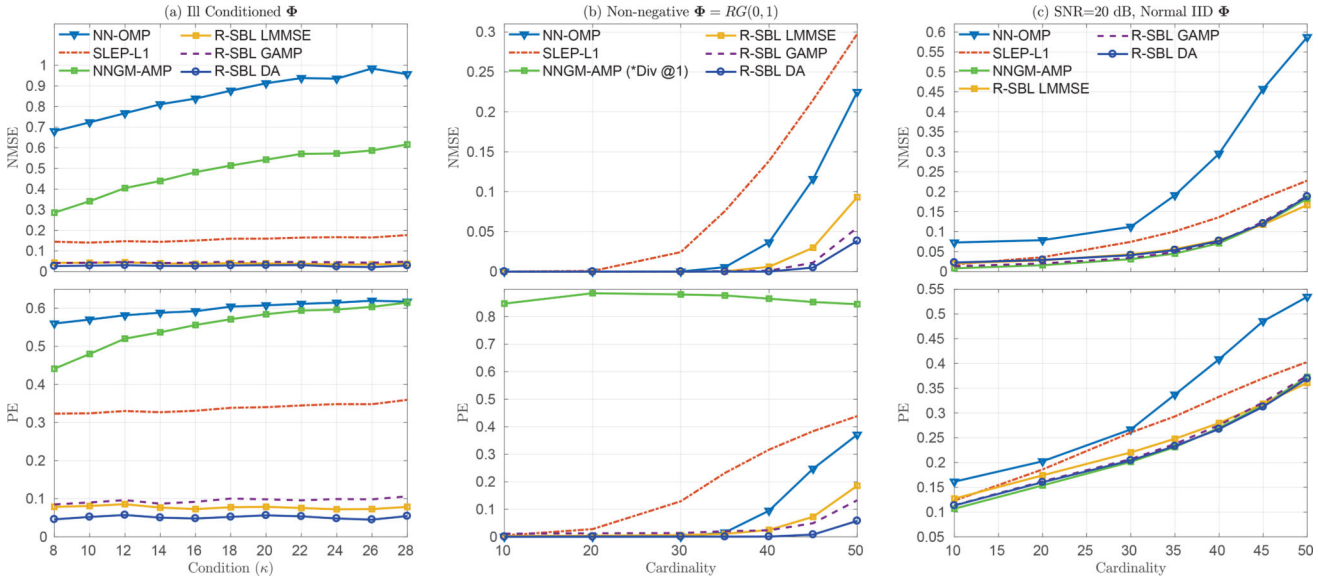

Fig. 3.

Sparse recovery performances of the S-NNLS solvers for various **Φ**. In (a), the dictionary is ill-conditioned with condition number κ given in the x-axis. R-SBL variants outperform the baseline solvers for various κ and are very robust to the value of κ. R-SBL DA achieves the lowest NMSE and PE. SLEP-ℓ_1 is superior to NNGM-AMP. In (b), the dictionary is non-negative with elements drawn from i.i.d. RG(0, 1). The recovery performances are given for various cardinality K in the x-axis. R-SBL variants achieve superior recovery across all values of K. NNGM-AMP diverges regardless of the value of K and is unable to recover a feasible solution. In (c), the dictionary is i.i.d. Normal and SNR is 20 dB. In the noisy setting, when the dictionary is i.i.d. and signed, R-SBL variants perform similar to NNGM-AMP under noisy conditions, but are superior to SLEP-ℓ_1 and NN-OMP at larger cardinalities.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

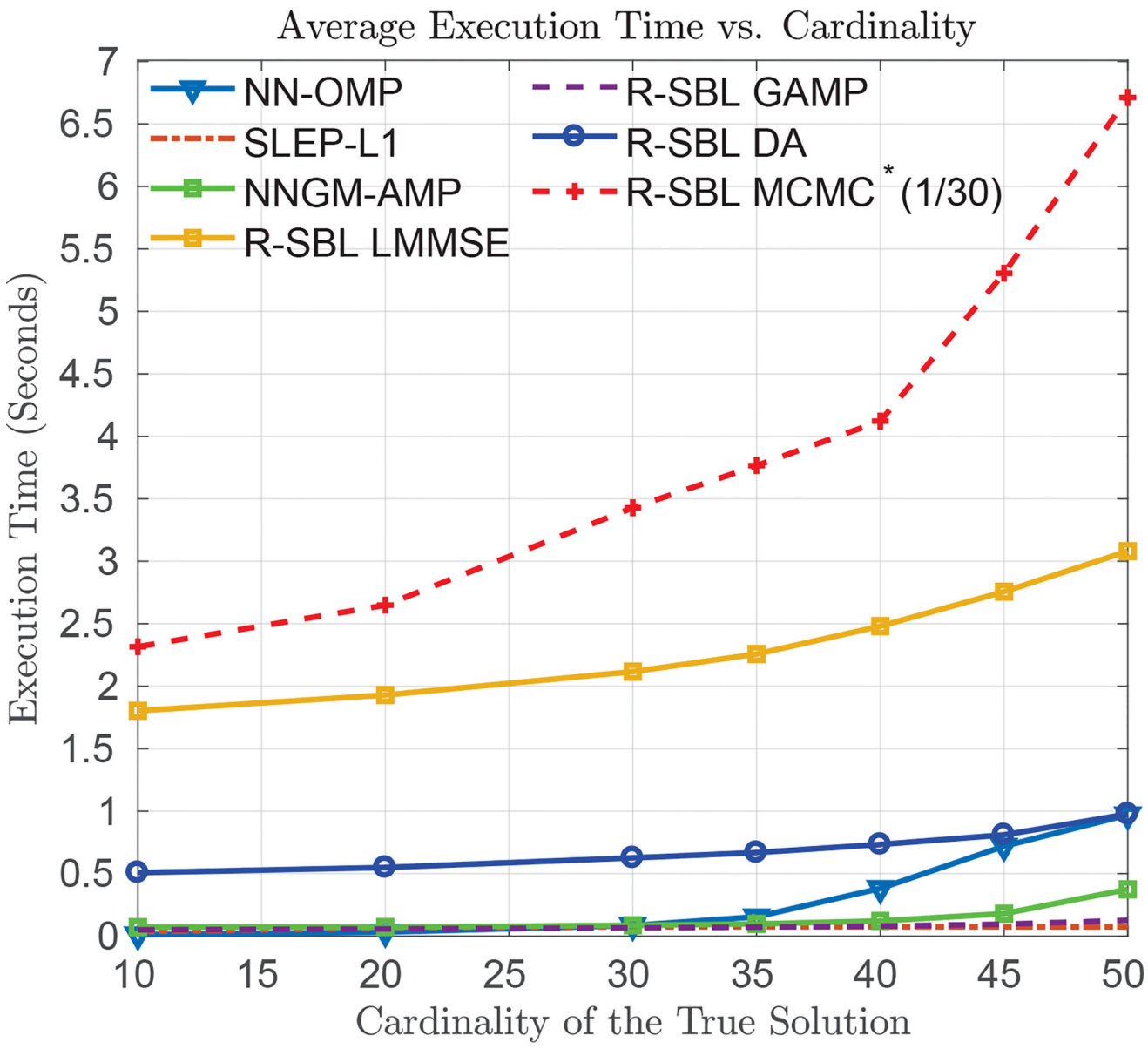

Fig. 4.
Execution times of the S-NNLS solvers as a function of cardinality for the noiseless scenario.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

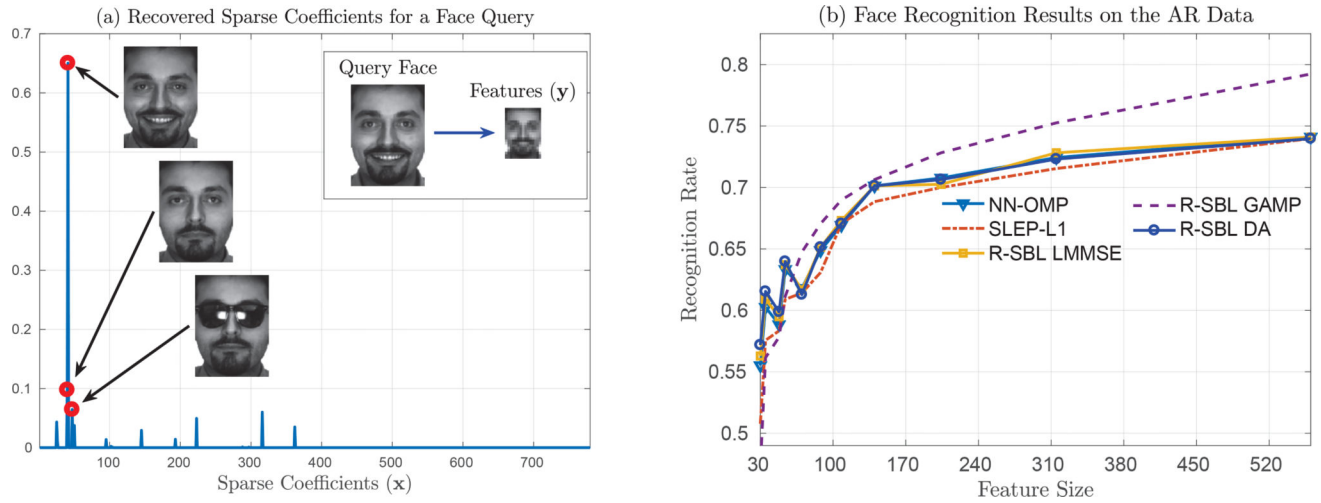

Fig. 5.

(a) Illustration of the sparse FR process. A query face is down-sample to obtain an observation $\boldsymbol{y}$. Using the training dictionary $\boldsymbol{\Phi}$, a sparse solution is obtained using R-SBL variants and baseline solvers to satisfy $\boldsymbol{y} = \boldsymbol{\Phi x}$. The index that corresponds to the maximum value in $\boldsymbol{x}$ is used to select a corresponding column from $\boldsymbol{\Phi}$. This corresponds to the correct individual. (b) FR accuracy for different feature sizes using all test samples. R-SBL GAMP enjoys better FR performance for different feature sizes.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE I

R-SBL GAMP Algorithm

Initialization

$\boldsymbol{S} \leftarrow |\boldsymbol{\Phi}|^2$ (component wise magnitude squared)

Initialize $\dot{\tau}_x^0 > 0$

$\boldsymbol{s}^0, \boldsymbol{x}^0 \leftarrow \mathbf{0}$

---

for $n = 1, 2, \ldots, N_{\max}$

Initialize $\boldsymbol{\tau}_x^1 \leftarrow \dot{\tau}_x^{n-1}, \boldsymbol{x}^1 \leftarrow \dot{\boldsymbol{x}}^{n-1}$

E-Step approximation

for $k = 1, 2, \ldots, K_{\max}$

$1/\boldsymbol{\tau}_p^k \leftarrow \boldsymbol{S}\boldsymbol{\tau}_x^k$

$\boldsymbol{p}^k \leftarrow \boldsymbol{s}^{k-1} + \boldsymbol{\tau}_p^k \boldsymbol{\Phi}\boldsymbol{x}^k$

$\boldsymbol{\tau}_s^k \leftarrow \frac{\sigma^{-2}\boldsymbol{\tau}_p^k}{\sigma^{-2} + \boldsymbol{\tau}_p^k}$

$\boldsymbol{s}^k \leftarrow (1-\theta_s)\boldsymbol{s}^{k-1} + \theta_s(\boldsymbol{p}^k/\boldsymbol{\tau}_p^k - \boldsymbol{y})/(\sigma^2 + 1/\boldsymbol{\tau}_p^k)$

$1/\boldsymbol{\tau}_r^k \leftarrow \boldsymbol{S}^\top \boldsymbol{\tau}_s^k$

$\boldsymbol{r}^k \leftarrow \boldsymbol{x}^k - \boldsymbol{\tau}_r^k \boldsymbol{\Phi}^\top \boldsymbol{s}^k$

if MaxSum then

$\boldsymbol{\tau}_x^{k+1} \leftarrow \boldsymbol{\nu}^k$

$\hat{\boldsymbol{x}}^{k+1} \leftarrow \boldsymbol{\eta}^k u(\boldsymbol{r}^k)$

else

$\boldsymbol{\tau}_x^{k+1} \leftarrow \boldsymbol{\nu}^k g(\frac{\boldsymbol{\eta}^k}{\boldsymbol{\nu}^k})$

$\hat{\boldsymbol{x}}^{k+1} \leftarrow \boldsymbol{\eta}^k + \sqrt{\boldsymbol{\nu}^k} h(\frac{\boldsymbol{\eta}^k}{\boldsymbol{\nu}^k})$

end if

if $\|\hat{\boldsymbol{x}}^{k+1} - \hat{\boldsymbol{x}}^k\|^2 / \hat{\boldsymbol{x}}^{k+1}\|^2 < \varepsilon_{\text{gamp}}$, break

end for %end of k loop

if MaxSum

$\dot{\boldsymbol{x}}^i \leftarrow \boldsymbol{\eta}^{k+1} + \sqrt{\boldsymbol{\nu}^{k+1}} h(\frac{\boldsymbol{\eta}^{k+1}}{\boldsymbol{\nu}^{k+1}}), \dot{\tau}_x^i \leftarrow \boldsymbol{\nu}^{k+1} g(\frac{\boldsymbol{\eta}^{k+1}}{\boldsymbol{\nu}^{k+1}})$

else

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

$\dot{\boldsymbol{x}}^n \leftarrow \hat{\boldsymbol{x}}^k + 1, \dot{\tau}_x^n \leftarrow \tau_x^k + 1$

end if

M-Step

$\gamma^n + 1 \leftarrow |\dot{\boldsymbol{x}}^n|^2 + \dot{\tau}_x^n$

if $\|\dot{\boldsymbol{x}}^n - \dot{\boldsymbol{x}}^{n-1}\|^2 / \|\dot{\boldsymbol{x}}n\|^2 < \epsilon_{\text{em}}$, break

end for %end of i loop

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE II

NMSE and PE results for various distributions for x^{gen}. The dictionary is i.i.d Normal distributed.

		Φ is i.i.d Normal					
	x^{gen}	**NN-OMP**	**SLEP-ℓ_1**	**NNGM AMP**	**R-SBL (LMMSE)**	**R-SBL (GAMP)**	**R-SBL (DA)**
NMSE	RG	0.4460	0.1439	0.0389	0.0488	0.0428	**0.0313**
	NN-Cauchy	0.0097	0.0086	0.0020	0.0004	0.0003	**0.0002**
	NN-Laplace	0.1566	0.0693	0.0091	0.0066	0.0059	**0.0034**
	Gamma	0.1476	0.0661	0.0074	0.0065	0.0045	**0.0024**
	Chi-square	0.1583	0.0673	0.0091	0.0077	0.0066	**0.0035**
	Bernoulli	0.5845	0.1265	**0.0052**	0.0524	0.0416	0.0339
PE	RG	0.4601	0.3208	0.0711	0.0873	0.0823	**0.0549**
	NN-Cauchy	0.2307	0.3509	0.2142	**0.0187**	0.0200	0.0408
	NN-Laplace	0.3202	0.3137	0.0407	0.0292	0.0229	**0.0118**
	Gamma	0.3091	0.3093	0.0416	0.0260	0.0207	**0.0080**
	Chi-square	0.3200	0.3086	0.0473	0.0307	0.0280	**0.0133**
	Bernoulli	0.4852	0.3283	**0.0101**	0.1714	0.1514	0.1264

IEEE Trans Signal Process. Author manuscript; available in PMC 2021 June 28.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE III

NMSE and PE results for various distributions for x^{gen}. The dictionary is i.i.d ±1 Bernoulli distributed.

		Φ is ±1 Bernoulli					
	x^{gen}	**NN-OMP**	**SLEP-ℓ_1**	**NNGM AMP**	**R-SBL (LMMSE)**	**R-SBL (GAMP)**	**R-SBL (DA)**
NMSE	RG	0.3996	0.1387	0.0409	0.0504	0.0415	**0.0332**
	NN-Cauchy	0.0083	0.0077	0.0023	0.0005	0.0004	**0.0003**
	NN-Laplace	0.1368	0.0712	0.0101	0.0096	0.0090	**0.0050**
	Gamma	0.1294	0.0665	0.0079	0.0061	0.0051	**0.0023**
	Chi-square	0.1267	0.0667	0.0109	0.0083	0.0094	**0.0055**
	Bernoulli	0.5610	0.1180	**0.0113**	0.0466	0.0412	0.0363
PE	RG	0.4272	0.3182	0.0794	0.0950	0.0824	**0.0568**
	NN-Cauchy	0.1810	0.3475	0.2307	0.0187	**0.0175**	0.0321
	NN-Laplace	0.2909	0.3131	0.0508	0.0369	0.0333	**0.0163**
	Gamma	0.2682	0.3072	0.0472	0.0274	0.0248	**0.0093**
	Chi-square	0.2769	0.3104	0.0532	0.0357	0.0369	**0.0195**
	Bernoulli	0.4734	0.3290	**0.0154**	0.1727	0.1571	0.1345

IEEE Trans Signal Process. Author manuscript; available in PMC 2021 June 28.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

TABLE IV

NMSE and PE results for various distributions for x^{gen}. The dictionary is i.i.d {0, 1} Bernoulli distributed.

		Φ is {0, 1} Bernoulli					
	x^{gen}	**NN-OMP**	**SEEP-E**	**NNGM AMP**	**R-SBL (LMMSE)**	**R-SBL (GAMP)**	**R-SBL (DA)**
NMSE	RG	0.2063	0.2497	Diverged	0.0873	0.0520	**0.0386**
	NN-Cauchy	0.0085	0.0188	Diverged	0.0031	0.0286	**0.0002**
	NN-Laplace	0.0960	0.1406	Diverged	0.0296	0.0070	**0.0043**
	Gamma	0.0901	0.1335	Diverged	0.0283	0.0047	**0.0022**
	Chi-square	0.0894	0.1360	Diverged	0.0327	0.0077	**0.0054**
	Bernoulli	0.2203	0.2586	Diverged	0.0747	0.0682	**0.0558**
PE	RG	0.3558	0.4070	0.8404	0.1782	0.0950	**0.0581**
	NN-Cauchy	0.2651	0.4434	0.8314	0.1131	0.4480	**0.0354**
	NN-Laplace	0.3140	0.4071	0.8398	0.1193	0.0371	**0.0134**
	Gamma	0.3102	0.4016	0.8354	0.1240	0.0275	**0.0087**
	Chi-square	0.3126	0.4072	0.8377	0.1341	0.0363	**0.0171**
	Bernoulli	0.3803	0.4120	0.8399	0.2689	0.1705	**0.1455**

IEEE Trans Signal Process. Author manuscript; available in PMC 2021 June 28.