POLITECNICO DI TORINO
Repository ISTITUZIONALE

Compressed sensing reconstruction using Expectation Propagation

Original
Compressed sensing reconstruction using Expectation Propagation / Braunstein, A., Muntoni, A.P., Pagnani, A., Pieropan, M.. - In: JOURNAL OF PHYSICS. A, MATHEMATICAL AND THEORETICAL. - ISSN 1751-8113. - 53:18(2020), p. 184001. [10.1088/1751-8121/ab3065]

Availability:
This version is available at: 11583/2741576 since: 2019-07-11T10:36:37Z

Publisher:
IOP Publishing

Published
DOI:10.1088/1751-8121/ab3065

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
IOP postprint/Author's Accepted Manuscript

“This is the accepted manuscript version of an article accepted for publication in JOURNAL OF PHYSICS. A, MATHEMATICAL AND THEORETICAL. IOP Publishing Ltd is not responsible for any errors or omissions in this version of the manuscript or any version derived from it. The Version of Record is available online at http://dx.doi.org/10.1088/1751-8121/ab3065

(Article begins on next page)

11 November 2026

Journal of Physics A: Mathematical and Theoretical

ACCEPTED MANUSCRIPT

Compressed sensing reconstruction using Expectation Propagation

To cite this article before publication: Alfredo Braunstein *et al* 2019 *J. Phys. A: Math. Theor.* in press https://doi.org/10.1088/1751-8121/ab3065

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2019 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.
As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 130.192.10.158 on 11/07/2019 at 09:23

Compressed sensing reconstruction using Expectation Propagation

Alfredo Braunstein[1,2,3,4], Anna Paola Muntoni[1,5,6], Andrea Pagnani[1,2,4] and Mirko Pieropan[1]‡

[1] Department of Applied Science and Technology (DISAT), Politecnico di Torino, Corso Duca degli Abruzzi 24, Torino, Italy

[2] Italian Institute for Genomic Medicine (IIGM) (former HuGeF), Via Nizza 52, Torino, Italy

[3] Collegio Carlo Alberto, Via Real Collegio 30, Moncalieri, Italy

[4] INFN Sezione di Torino, Via P. Giuria 1, Torino, Italy

[5] Laboratoire de Physique de l'Ecole Normale Supérieure, ENS, Université PSL, CNRS, Sorbonne Université, Université de Paris, Paris, France

[6] Sorbonne Université, CNRS, Institut de Biologie Paris-Seine, Laboratory of Computational and Quantitative Biology, F-75005, Paris, France

E-mail: `mirko.pieropan@polito.it`

Abstract.

Many interesting problems in fields ranging from telecommunications to computational biology can be formalized in terms of large underdetermined systems of linear equations with additional constraints or regularizers. One of the most studied ones, the Compressed Sensing problem (CS), consists in finding the solution with the smallest number of non-zero components of a given system of linear equations $\boldsymbol{y} = \mathbf{F}\boldsymbol{w}$ for known measurement vector $\boldsymbol{y}$ and sensing matrix $\mathbf{F}$. Here, we will address the compressed sensing problem within a Bayesian inference framework where the sparsity constraint is remapped into a singular prior distribution (called Spike-and-Slab or Bernoulli-Gauss). Solution to the problem is attempted through the computation of marginal distributions via Expectation Propagation (EP), an iterative computational scheme originally developed in Statistical Physics. We will show that this strategy is more accurate for statistically correlated measurement matrices. For computational strategies based on the Bayesian framework such as variants of Belief Propagation, this is to be expected, as they implicitly rely on the hypothesis of statistical independence among the entries of the sensing matrix. Perhaps surprisingly, the method outperforms uniformly also all the other state-of-the-art methods in our tests.

‡ Author to whom any correspondence should be addressed

Accepted Manuscript

1. Introduction

The problem of Compressed Sensing (CS) [1, 2, 3] has led to significant developments in the field of sparse approximation and representation [4], together with the more traditional framework of optimal signal processing [4, 5].

In general, CS deals with the reconstruction of a sparse $N-$dimensional signal from M, often noisy, measurements. In this context, a sparse signal is characterized by a large (possibly the largest) number of components equal to zero. In practical cases, the regime of interest is given by the $M \ll N$, $K \ll N$ limit of the problem, where K is the number of non-zero components. A brute-force exhaustive search of the $K-$dimensional minimal support of the $N-$dimensional signal would lead to an exponential explosion in the search space as there are $\binom{N}{K}$ possible base supports to be explored. In the noiseless limit of CS, even if the K-dimensional minimal base were given, at least $M = K$ measurements are necessary to uniquely identify the solution.

CS can be easily formulated: let $\boldsymbol{w} \in \mathbb{R}^N$ be compressed into a vector $\boldsymbol{y} \in \mathbb{R}^M$ (from now on we will assume $M < N$) through a linear transformation:

$$\boldsymbol{y} = \mathbf{F}\boldsymbol{w}, \tag{1}$$

where $\mathbf{F} \in \mathbb{R}^{M\times N}$ is a linear operator of maximal rank often referred to as the *measurement* or *sensing* matrix. The problem is to determine the vector $\boldsymbol{w}$ from the knowledge of the measurement matrix $\mathbf{F}$ and of the compressed vector $\boldsymbol{y}$.

The relevance of CS in statistical physics stems from the link with the statistical mechanics of disordered systems, in analogy with many other combinatorial optimization problems [6, 7, 8]. In particular CS is an optimization problem with quenched disorder (the measurement matrix) and, as such, it is amenable to analytic treatment using replica theory [9] as developed in the field of spin-glasses [10].

From equation (1), it is clear that, as long as $M < N$, there are infinitely many solutions to the system of equations. However, if one imposes the supplementary condition on the sparsity of the signal $\boldsymbol{w}$, solving the problem may still lead to a unique retrieved signal. Thus, a problem in CS is to determine the set of conditions under which it is possible to find a solution of equation (1) that is as sparse as possible. A practical way to enforce sparsity is through the minimization of the L_p-norm of the $\boldsymbol{w}$ vector in the space of solutions which is defined in equation (1). A particularly successful line of research has been pursued through the minimization of the L_1 norm [1, 2, 3, 9, 11, 12, 13], for which the following prediction was obtained in the case of independent and identically distributed (i.i.d.) random Gaussian measurement matrices: in the large N limit, there is a non-trivial region $\rho = K/N$, $\alpha = M/N$ where an exact reconstruction of the original signal is indeed possible. The parameters ρ and α that characterize the signal are called the *signal density* and the *measurement rate*, respectively.

However, in many applications, the sensing matrix may not be random [14] or the i.i.d. assumption might not hold and, in these cases, several algorithms could fail to decode the signal. Examples of deterministic matrices include chirp sensing matrices

Accepted Manuscript

[15], which have been applied to image reconstruction [16], and second-order Reed-Muller matrices [17], whereas examples of correlated random matrices include random partial Fourier matrices [18, 19], which are encountered in MRI [20] as well as in other applications [21, 22], and partial random circulant and Toeplitz matrices [23], which arise in the presence of convolutions.

Our approach here, is to treat CS as a Bayesian inference problem, i.e. we focus our attention on determining the probability of observing, *a posteriori*, the signal $\boldsymbol{w}$ when we have observed a set of measurements $\boldsymbol{y}$ and given the sensing matrix $\mathbf{F}$. While equation (1) is easily cast in a Gaussian-like likelihood function, the prior over the variables $\boldsymbol{w}$, enforcing the sparsity constraint, plays a key role on the goodness of the solution and on the complexity of the problem. A prior that corresponds to the minimization of a L_p norm, with $p = 1, 2$, allows an easy marginalization of the corresponding posterior probability with the drawback of softening the sparsity requirement. Similarly to [9, 24, 25], we consider the so-called spike-and-slab prior [26] which is understood as the minimization of the L_0 norm of the vector $\boldsymbol{w}$. The posterior distribution results to be intractable in practice and thus some approximations need to be sought.

Almost twenty years ago Expectation Propagation (EP), an iterative scheme to approximate intractable distributions, has been introduced first in the field of statistical physics [27, 28] and shortly after in the field of theoretical computer science [29]. Recently, EP inspired inference strategies – similar to the one we present here – have been proposed to solve other underdetermined linear constraint problems such as the problem of sampling solutions from the reconstruction of large scale metabolic networks [30] and of tomographic images [31]. Here we propose an efficient and accurate EP-based reconstruction strategy for CS which, moreover, does not require i.i.d. measurement matrices.

Other attempts to solve sparse linear models using the EP approximation can be found in [32] – where the authors use a similar EP implementation to the one we adopt here and make use of a Laplace prior – and in [33] – where, in addition to the spike-and-slab, a Bernoulli prior on the components of the signal is introduced and the EP update scheme involves only three approximating factors of the original posterior distribution.

All these approaches have the same computational complexity, which is dominated by a single matrix inversion per EP cycle. In our implementation, we integrate the inference of the signal with a variational learning of the density parameter of the spike-and-slab prior. Besides, we show in Appendix A that the original EP scheme is equivalent to an alternative formulation of the EP update equations that takes into account the linear constraints (1) exactly. This implementation has the advantage of reducing the size of the matrix to be inverted, thus decreasing the computing time. Throughout the work presented in this paper, we have numerically checked that both procedures can be used interchangeably in the case of noiseless measurements, as they lead to the same results.

In conclusion, we believe that our EP algorithm shows some original features that have not been proposed by other methods, namely: (i) the quality of EP

Accepted Manuscript

reconstruction on correlated measurement matrices is the same as in the case of uncorrelated measurements whereas all other methods we tested fail; (ii) to reconstruct signals in the noiseless case, we introduced an EP formulation performing *analytically* the zero temperature limit; (iii) we learn the density parameter of the spike and slab prior from the data by minimizing iteratively the EP free energy.

The outline is as follows: we introduce the EP approach to CS and set the basic notations in section 2, we show a thorough comparison with other state of the art algorithms in section 3 and, finally, we present our conclusions in section 4.

2. Expectation Propagation

In this section we introduce the so-called finite temperature formulation of EP, where we allow the measurement vector $\boldsymbol{y}$ to be noisy. We consider the *undersampling* regime $M < N$, although there is no technical limitation in considering the $M \geq N$ case. The linear constraints in equation (1) can be alternatively mapped into a minimization problem of the following quadratic form:

$$E(\boldsymbol{w}) = \frac{1}{2} \left\| \boldsymbol{y} - \mathbf{F}\boldsymbol{w} \right\|_2 := \frac{1}{2} (\boldsymbol{y} - \mathbf{F}\boldsymbol{w})^T \cdot (\boldsymbol{y} - \mathbf{F}\boldsymbol{w}) \,, \tag{2}$$

From a Bayesian perspective, the probability of observing the vector $\boldsymbol{y}$ given the matrix $\mathbf{F}$ and the vector $\boldsymbol{w}$ is:

$$P(\boldsymbol{y}|\mathbf{F}, \boldsymbol{w}) = \left(\frac{\beta}{2\pi} \right)^{\frac{M}{2}} e^{-\beta E(\boldsymbol{w})}, \tag{3}$$

where we introduced a fictitious inverse temperature β that we can take as large as we wish in order to enforce the linear bounds expressed by equation (1). Alternatively we can interpret equation (3) as the probability of observing an additive noise vector $\boldsymbol{y} - \mathbf{F}\boldsymbol{w}$ whose elements are distributed according to a Gaussian density of zero mean and variance β^{-1}. The posterior distribution for the vector $\boldsymbol{w}$ reads:

$$P\left(\boldsymbol{w}|\mathbf{F}, \boldsymbol{y}\right) = \frac{P\left(\boldsymbol{y}|\mathbf{F}, \boldsymbol{w}\right) P\left(\boldsymbol{w}\right)}{P\left(\boldsymbol{y}\right)} = \frac{1}{Z_P} e^{-\beta \frac{(\boldsymbol{y} - \mathbf{F}\boldsymbol{w})^T \cdot (\boldsymbol{y} - \mathbf{F}\boldsymbol{w})}{2}} \prod_{i=1}^{N} \psi_i(w_i), \tag{4}$$

where in the last step we restricted the structure of the prior $P(\boldsymbol{w}) := \prod_{i=1}^{N} \psi_i(w_i)$ to a factorized form, although more general structures can be considered, e.g. as in [34]. In this work, we have considered the so-called spike-and-slab prior [26]:

$$\psi(w_i) = (1 - \rho)\delta(w_i) + \frac{\rho}{\sqrt{2\pi\lambda}} e^{-\frac{w_i^2}{2\lambda}}, \tag{5}$$

where δ denotes the Dirac delta function, in order to model any prior knowledge about the sparsity of the signal.

We seek a solution vector $\hat{\boldsymbol{w}}$ whose components are the first moments of the marginal densities of equation (4). Contrarily to the maximum a posteriori estimate, it can be proven that this strategy minimizes the mean squared error between the true and the recovered signal. Unfortunately, due to the non-convex nature of the spike-and-slab prior, there exists no technique able to perform the marginalization of the posterior

Accepted Manuscript

probability in equation (4). In the following we introduce the EP approximation scheme which relies on an adaptive Gaussian approximation of the marginal probabilities of interest.

2.1. The approximate posterior distribution

EP [29] is an efficient approximation to compute posterior probabilities. EP was first introduced as an improved mean-field method [27, 28] and further developed in the framework of Bayesian inference problems in the seminal work of Minka [29]. The approximated distribution consists in substituting the typically analytically intractable ψ_i priors with univariate Gaussian distributions $\phi_i(w_i) = \mathcal{N}(w_i; a_i, d_i)$ of mean a_i and variance d_i. The approximated posterior thus reads:

$$\begin{aligned} Q(\boldsymbol{w}|\mathbf{F},\boldsymbol{y}) = & \frac{1}{Z_Q} e^{-\beta \frac{(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})^T\cdot(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})}{2}} \prod_{i=1}^{N} \phi_i(w_i) \\ := & \frac{1}{Z_Q} e^{-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{w}-\bar{\boldsymbol{w}})}, \end{aligned} \tag{6}$$

where:

$$Z_Q = (2\pi)^{\frac{N}{2}} (\det \boldsymbol{\Sigma})^{\frac{1}{2}}, \tag{7}$$

$$\boldsymbol{\Sigma}^{-1} := \beta \mathbf{F}^T\mathbf{F} + \boldsymbol{D}, \quad \bar{\boldsymbol{w}} := \boldsymbol{\Sigma}(\beta \mathbf{F}^T \boldsymbol{y} + \boldsymbol{D}\boldsymbol{a}), \tag{8}$$

and $\boldsymbol{D}$ is a diagonal matrix having diagonal elements $d_1^{-1}, \ldots, d_N^{-1}$. We now need a way to fix the parameters $\boldsymbol{a}, \boldsymbol{d}$ of the prior. To do so, we focus on the n^{th} variable w_n (with $1 \leq n \leq N$), and in particular on its approximated prior ϕ_n. We can define the *tilted* distribution $Q^{(n)}$ as:

$$\begin{aligned} Q^{(n)}(\boldsymbol{w}|\mathbf{F},\boldsymbol{y}) := & \frac{1}{Z_{Q^{(n)}}} e^{-\beta \frac{(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})^T\cdot(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})}{2}} \psi_n(w_n) \prod_{l\neq n} \phi_l(w_l; a_l, d_l) \\ = & \frac{1}{Z_{Q^{(n)}}} e^{-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}}_{(n)})^T \Sigma_{(n)}^{-1} (\boldsymbol{w}-\bar{\boldsymbol{w}}_{(n)})} \psi_n(w_n), \end{aligned} \tag{9}$$

where:

$$\boldsymbol{\Sigma}_{(n)}^{-1} = \beta \mathbf{F}^T\mathbf{F} + \boldsymbol{D}_{(n)}, \quad \bar{\boldsymbol{w}}_{(n)} = \boldsymbol{\Sigma}_{(n)} \left(\beta \mathbf{F}^T \boldsymbol{y} + \boldsymbol{D}_{(n)}\boldsymbol{a}\right), \tag{10}$$

and, in analogy with equation (8), $\boldsymbol{D}_{(n)}$ is a diagonal matrix of elements d_m^{-1} for all diagonal elements $m \neq n$ and zero for $m = n$. The *tilted* distribution differs from the approximated posterior Q as it contains the original prior ψ_n instead of the n-th approximated prior ϕ_n. Intuitively, we expect that the n-th *tilted* distribution provides a better approximation of the expectation values related to the n-th variable than the multivariate Gaussian approximation. From a computational point of view, the presence of a single intractable prior in the *tilted* distribution does not prejudice the efficiency of the algorithm.

A natural way of fixing the optimal $\boldsymbol{a}, \boldsymbol{d}$ parameters consists in requiring the approximated Q distribution to be as similar as possible to the *tilted* distribution $Q^{(n)}$.

Accepted Manuscript

To do so, we minimize the Kullback-Leibler divergence $D_{KL}\left(Q^{(n)}||Q\right)$. Perhaps not surprisingly, this procedure is equivalent to equating the first two moments of the two distributions:

$$\langle w_n \rangle_{Q^{(n)}} = \langle w_n \rangle_Q , \quad \left\langle w_n^2 \right\rangle_{Q^{(n)}} = \left\langle w_n^2 \right\rangle_Q , \tag{11}$$

where $\langle \cdot \rangle_Q$ and $\langle \cdot \rangle_{Q^{(n)}}$ denote averages w.r.t. $Q(\boldsymbol{w})$ and $Q^{(n)}(\boldsymbol{w})$, respectively.

Notice that the computation of the moments of the *tilted* distribution on the left-hand side of equation (11) depends on the functional form of the prior considered. We refer to Appendix B for the expression of the moments of the tilted distributions used in the case of a spike-and-slab prior.

Thanks to the multivariate Gaussian form of the approximated distribution, it is a simple exercise to compute the moments of Q:

$$\begin{aligned} \langle w_n \rangle_Q &= \left(\frac{1}{d_n} + \frac{1}{\Sigma_{n,n}} \right)^{-1} \left(\frac{a_n}{d_n} + \frac{\bar{w}_n}{\Sigma_{n,n}} \right), \\ \langle w_n^2 \rangle_Q &= \frac{1}{\frac{1}{d_n} + \frac{1}{\Sigma_{n,n}}} + \langle w_n \rangle_Q^2 . \end{aligned} \tag{12}$$

From equations (8) and (10), it is clear that the two matrices Σ^{-1}, and $\Sigma_{(n)}^{-1}$ differ only in a diagonal term. We can thus exploit a low-rank update property to relate the two inverses. It turns out that the tilted parameters are related to the approximated ones:

$$\left(\bar{w}_{(n)}\right)_n = \left(\Sigma_{(n)}\right)_{n,n} \left(\frac{\bar{w}_n}{\Sigma_{n,n}} - \frac{a_n}{d_n} \right), \quad \left(\Sigma_{(n)}\right)_{n,n} = \frac{\Sigma_{n,n}}{1 - \frac{\Sigma_{n,n}}{d_n}} . \tag{13}$$

Upon imposing the moment matching condition (11), we eventually get an explicit equation for the prior parameters a_n, d_n:

$$\begin{aligned} d_n &= \left(\frac{1}{\langle w_n^2 \rangle_{Q^{(n)}} - \langle w_n \rangle_{Q^{(n)}}^2} - \frac{1}{\left(\Sigma_{(n)}\right)_{n,n}} \right)^{-1} , \\ a_n &= \langle w_n \rangle_{Q^{(n)}} + \frac{d_n}{\left(\Sigma_{(n)}\right)_{n,n}} \left(\langle w_n \rangle_{Q^{(n)}} - \left(\bar{w}_{(n)}\right)_n \right) . \end{aligned} \tag{14}$$

The $\boldsymbol{a}, \boldsymbol{d}$ parameters are sequentially updated until a fixed point is eventually reached: numerically, we need to set a threshold below which the algorithm stops. To this purpose, for each iteration t (i.e. for each update of the $\boldsymbol{a}, \boldsymbol{d}$ vectors), we can define an error ϵ_t as:

$$\epsilon_t = \max_n \left| \langle w_n \rangle_{Q_t^{(n)}} - \langle w_n \rangle_{Q_{t-1}^{(n)}} \right| + \left| \left\langle w_n^2 \right\rangle_{Q_t^{(n)}} - \left\langle w_n^2 \right\rangle_{Q_{t-1}^{(n)}} \right| ,$$

where $Q_t^{(n)}$ is the tilted distribution with parameters computed at iteration t. In practice, the algorithm stops when $\epsilon_t < 10^{-6}$. At convergence, the tilted distributions provide an approximation to the marginal densities of the posterior in equation (4) and their first moments $\langle w_n \rangle_{Q^{(n)}}$ provide the estimate $\hat{\boldsymbol{w}}$ of the unknown vector $\boldsymbol{w}$.

For the sake of convenience, the EP procedure with low rank update that we have just presented is summarized in Algorithm 1.

Accepted Manuscript

Algorithm 1 Expectation Propagation with low rank update

procedure EP($\mathbf{F}$, $\boldsymbol{y}$,$\{\psi_1, ..., \psi_N\}$)
  Initialize $\boldsymbol{a}^{old}$ and $\boldsymbol{d}^{old}$
  $\mathbf{A} = \beta \mathbf{F}^T \mathbf{F}$
  repeat
    $\boldsymbol{\Sigma} = (\mathbf{A} + \mathbf{D})^{-1}$
    $\bar{\boldsymbol{w}} = \boldsymbol{\Sigma}(\beta \mathbf{F}^T \boldsymbol{y} + \mathbf{D}\boldsymbol{a})$
    for $k = 1, ..., N$ **do**
      Compute $\mu_k^{(k)}$ and $\Sigma_{kk}^{(k)}$ using the low rank update (13).
      Compute moments $\langle x_k \rangle_{Q^{(k)}}$ and $\langle x_k^2 \rangle_{Q^{(k)}}$.
      Compute d_k^{new} and a_k^{new} by moment matching using equation (14).
  until convergence
  return averages $\left\{ \langle x_k \rangle_{Q^{(k)}} \right\}_{k=1,...,N}$ and variances $\left\{ \langle x_k^2 \rangle_{Q^{(k)}} - \langle x_k \rangle_{Q^{(k)}}^2 \right\}_{k=1,...,N}$.

3. Experimental results with synthetic data

In this section, we present our empirical results obtained by means of numerical simulations. We will first consider compressed sensing with i.i.d. random sensing matrices and then we will discuss some results related to the case of correlated random matrices.

3.1. Uncorrelated measurements

We consider $M \times N$ measurement matrices having i.i.d. entries sampled from a standard normal distribution $\mathcal{N}(0, 1)$. The signal vector has $K = \rho N$ nonzero components, which are also sampled from a Gaussian distribution with zero mean and unit variance. The measurements are assumed to be noiseless. Note that, in general, the parameter ρ in equation (5) is unknown and needs to be estimated. We show in Appendix C how one can infer ρ within the framework provided by EP.

We have run EP throughout the ρ-α plane. The parameters used in our EP simulations are $\lambda = 1$ and, in the finite temperature formulation, $\beta = 10^9$.

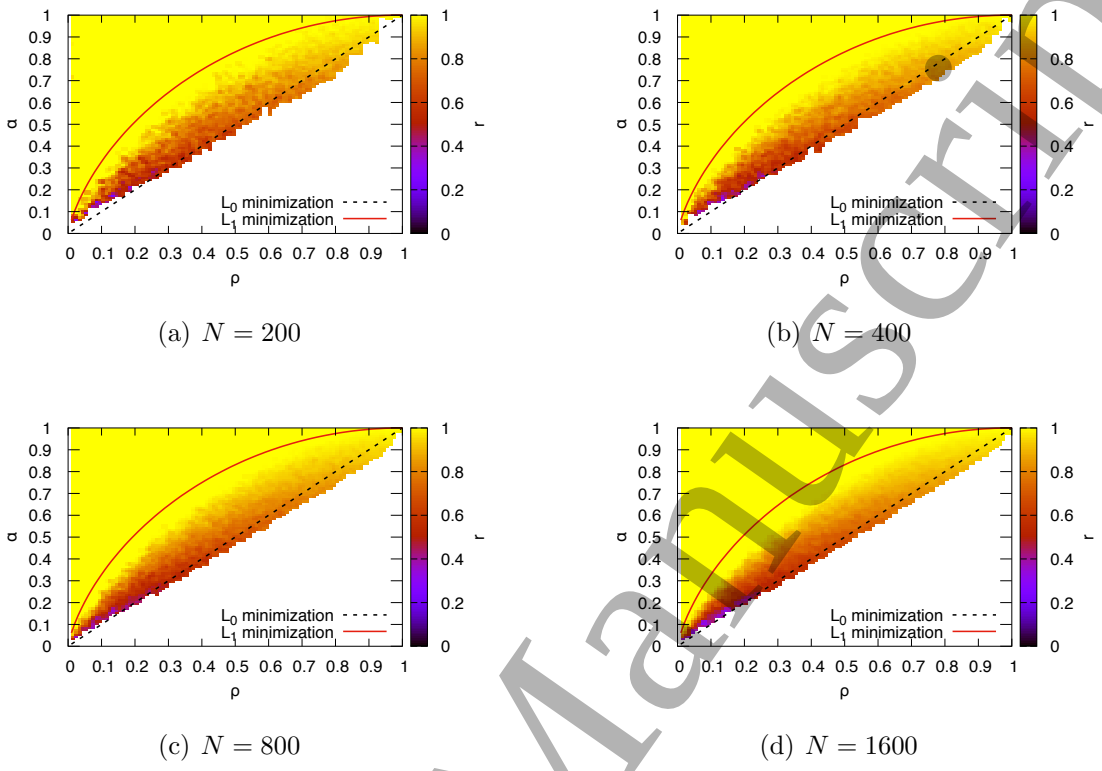
In order to measure the quality of the reconstruction, we consider the sample Pearson correlation coefficient r of the true vector and of the reconstructed vector:

$$r = \frac{\sum_{k=1}^{N} (w_k - w_{sm})(\hat{w}_k - \hat{w}_{sm})}{\sqrt{\sum_{k=1}^{N} (w_k - w_{sm})^2} \sqrt{\sum_{k=1}^{N} (\hat{w}_k - \hat{w}_{sm})^2}}, \tag{15}$$

where w_{sm} and $\hat{w}_{sm}$ are the sample means of the signal and of the inferred vector, respectively. We also consider the within-sample mean squared error as a measure of the reconstruction error:

$$MSE = \frac{1}{N} \sum_{k=1}^{N} (w_k - \hat{w}_k)^2. \tag{16}$$

Accepted Manuscript



(a) $N = 200$ (b) $N = 400$

(c) $N = 800$ (d) $N = 1600$

Figure 1. Compressed sensing phase diagram for $N = 200, 400, 800, 1600$. Each point corresponds to a single simulation. The color refers to the Pearson correlation coefficient between the true vector and the reconstructed vector. We plot the lines corresponding to the L_1 reconstruction and to the theoretical reconstruction limit, given by the L_0 condition.

We plot the sample correlation coefficient in figure 1 for a single simulation at each given ρ and α and progressively larger values of N, namely $N = 200$ (figure 1(a)), $N = 400$ (figure 1(b)), $N = 800$ (figure 1(c)) and $N = 1600$ (figure 1(d)). The L_0 line represents the theoretical limit $M = K$, under which a perfect reconstruction is impossible, whereas the L_1 line was obtained in [9, 35] using the replica method in the limit $N \to \infty$ and $M \to \infty$, with α finite. The white region is the one in which EP does not converge.

The plots suggest that there exists a phase transition line $\{(\rho, \alpha^{EP}(\rho)), 0 \leq \rho \leq 1\}$, which is located under the L_1 line. In order to obtain the coordinates of the points along the transition line, one can proceed numerically by using a bisection-like algorithm. After discretizing the interval $0 \leq \rho \leq 1$, one can select two starting values of α for each discretized value ρ_0. For instance, a reasonable choice, which is the one we adopt, is taking α_0 on the L_0 minimization line, namely $\alpha_0(\rho_0) = \rho_0$, and $\alpha_1(\rho_0)$ on the L_1 minimization line, where $\alpha_1(\rho_0)$ is expressed as [9]:

$$\alpha_1(\rho_0) = 2(1 - \rho_0)H(\hat{\chi}^{-1/2}) + \rho_0. \tag{17}$$

Accepted Manuscript

In the last equation, $\hat{\chi}$ is given by the solution of [35]:

$$\hat{\chi} = \alpha^{-1}\left[2(1-\rho_0)\left((\hat{\chi}+1)H(\hat{\chi}^{-1/2}) - \hat{\chi}^{1/2}\frac{e^{-1/(2\hat{\chi})}}{\sqrt{2\pi}}\right) + \rho_0(\hat{\chi}+1)\right], \quad (18)$$

where $H(x) = \int_x^{+\infty}\exp(-\frac{t^2}{2})/(2\pi)dt$. Then, one performs the EP inference for configurations corresponding to those points and to the point (ρ^*, α^*), where $\rho^* = \rho_0$ and $\alpha^* = (\alpha_0 + \alpha_1)/2$ and computes their mean squared error. If the difference $|MSE(\alpha_1) - MSE(\alpha^*)|$ is negligible (i. e. smaller than a certain threshold δ), then we set $\alpha_1 = \alpha^*$. Otherwise, we set $\alpha_0 = \alpha^*$. We recompute the middle point and repeat the procedure until we reach the desired accuracy $\Delta\alpha_{min}$ on the points located at the boundary between the successful and unsuccessful reconstruction regions. We summarize the procedure in Algorithm 2. The resulting transition line is shown in figure 2.

Algorithm 2 Bisection algorithm

procedure BISECTION($N, \rho_0, \alpha_0, \alpha_1; \delta, \Delta\alpha_{min}$)
- Set $K = \rho_0 N$.
- Set $\alpha^* = (\alpha_0 + \alpha_1)/2$.
- **repeat**
  - Set $M_1 = \alpha_1 N$.
  - Generate signal $\boldsymbol{w}_1$ and sensing matrix $\mathbf{F}_1$.
  - Infer $\hat{\boldsymbol{w}}_1$ using Algorithm 1 with inputs $\boldsymbol{y}_1 = \mathbf{F}_1\boldsymbol{w}_1$, $\mathbf{F}_1$ and $\rho = \rho_0$.
  - Compute $MSE(\alpha_1)$ between $\hat{\boldsymbol{w}}_1$ and $\boldsymbol{w}_1$.
  - Set $M^* = \alpha^* N$.
  - Generate signal $\boldsymbol{w}^*$ and sensing matrix $\mathbf{F}^*$.
  - Infer $\hat{\boldsymbol{w}}^*$ using Algorithm 1 with inputs $\boldsymbol{y}^* = \mathbf{F}^*\boldsymbol{w}^*$, $\mathbf{F}^*$ and $\rho = \rho_0$.
  - Compute $MSE(\alpha^*)$ between $\hat{\boldsymbol{w}}^*$ and $\boldsymbol{w}^*$.
  - **if** $|MSE(\alpha_1) - MSE(\alpha^*)| > \delta$ **then**
    - $\alpha_0 = \alpha^*$.
  - **else**
    - $\alpha_1 = \alpha^*$.
  - Reassign $\alpha^* = (\alpha_0 + \alpha_1)/2$.
- **until** $|\alpha_1 - \alpha_0|/2 < \Delta\alpha_{min}$
- **return** α^*

We define *probability of convergence* as the empirical frequency that for a random instance of the signal $\boldsymbol{w}$ and of the measurement matrix $\mathbf{F}$, the algorithmic error ϵ_t becomes arbitrarily small after some iteration, as the maximum number of iterations t is increased. In practice, we set a threshold for the error ϵ_t equal to 10^{-6} and we estimated the probability of convergence as the fraction of times the algorithm fulfilled the convergence criterion. Empirically, it turns out that the probability of convergence of EP, increases with the number of variables N (not shown). Moreover, the fluctuations of

Accepted Manuscript

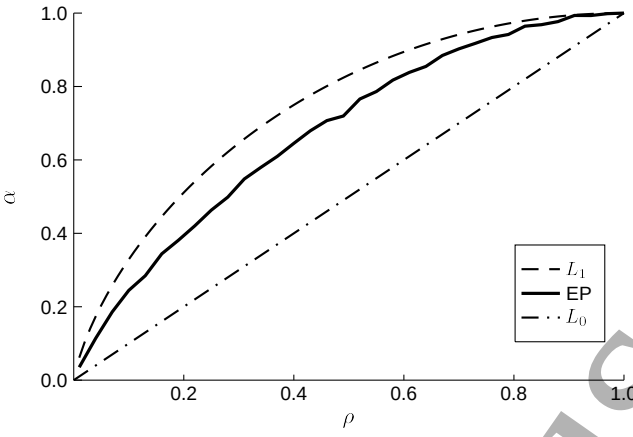
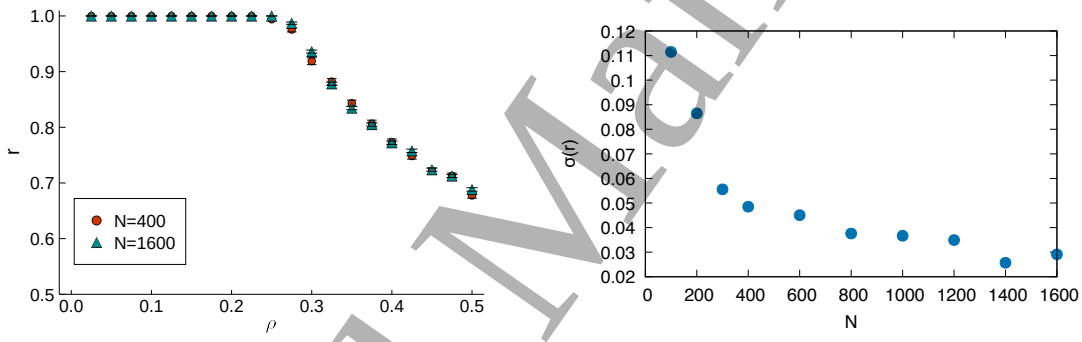


Figure 2. EP phase transition line as obtained using the bisection-like algorithm described in the main text. The number of variables is $N = 1600$ and the threshold δ for the difference of the mean squared error of the evaluated points is 10^{-5}.



(a) (b)

Figure 3. (a) Pearson correlation as a function of ρ at $\alpha = 0.5$, for $N = 400$ and $N = 1600$. The error bars are estimated as $\sigma(r)/N_t$, where $N_t = 100$ is the number of trials. (b) Sample standard deviation of r as a function of N. The value of the density and of the measurement rate are fixed and given by $\rho = 0.4$ and $\alpha = 0.55$.

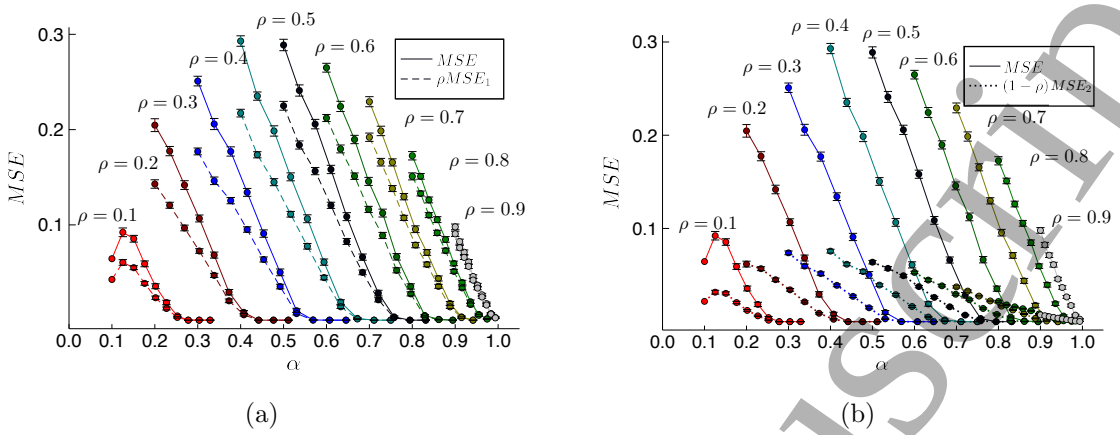
the Pearson correlation coefficient r and of the MSE beyond the transition line decrease as the number of variables N becomes larger, whereas their average values do not seem to depend on the size of the system. We show this in the case of the MSE in figure 3, for $N = 400$ and $N = 1600$.

Finally, we note that the mean squared error can be expressed as follows:

$$MSE = \rho MSE_1 + (1 - \rho)MSE_2, \tag{19}$$

where MSE_1 is the mean squared error between the vector of the K non zero components of the signal and the corresponding vector extracted from the inferred signal and MSE_2 is the mean squared error between the original and inferred vectors having the remaining $N - K$ elements as components. The latter corresponds to the squared norm of the last $N - K$ components of the inferred vector (divided by $N - K$). Beyond the CS threshold of EP, the dominant contribution to the reconstruction error comes from the estimate of the K non-zero components of the reconstructed vector, implying that, overall, EP

Accepted Manuscript



(a) (b)

Figure 4. a) Contribution of the first K components to the mean squared error (dashed lines), compared to the mean squared error itself (solid lines). (b) Contribution of the last $N-K$ components (the 'tail' of the vector) to the mean squared error (dotted lines), compared to the mean squared error itself (solid lines). In all plots, $N = 400$, the number of simulations is 100 and each curve corresponds to a different value of ρ. The points are averages computed over the N_c converged simulations and the error is estimated from the sample standard deviation σ as $\sigma/\sqrt{N_c}$. From left to right, ρ ranges from 0.1 to 0.9.

is still quite accurate in discriminating the zero entries of the signal. This is shown in figures 4(a) and 4(b), in which, respectively, ρMSE_1 and $(1-\rho)MSE_2$ are compared to the total mean squared error.

3.2. Correlated measurement matrices

We consider the case of correlated measurement matrices $\mathbf{F}$:

$$\mathbf{F} = (\underline{f}_1, ..., \underline{f}_M)^T, \tag{20}$$

whose rows $\underline{f}_i \in \mathbb{R}^N$ are correlated but linearly independent samples drawn from a multivariate Gaussian distribution:

$$\underline{f}_i \sim \mathcal{N}(\underline{0}, \mathbf{S}), \qquad i = 1, ..., M. \tag{21}$$

The covariance matrix $\mathbf{S}$ is designed according to the following functional form:

$$\mathbf{S} = \mathbf{Y}^T\mathbf{Y} + \mathbf{\Delta}, \tag{22}$$

where $\mathbf{Y}$ is a $k \times N$ matrix with random i.i.d. Gaussian $\mathcal{N}(0,1)$ entries and having a controllable rank k and $\mathbf{\Delta}$ is a diagonal $N \times N$ matrix with positive Gaussian i.i.d. eigenvalues. Notice that the product $\mathbf{Y}^T\mathbf{Y}$ is symmetric and positive semi-definite by construction. Adding the matrix $\mathbf{\Delta}$ ensures that $\mathbf{S}$ has maximum rank.

We first study the retrieval performance of EP and of Expectation Maximization Belief Propagation (EMBP), a similar message passing reconstruction algorithm [24, 25], implemented in MATLAB and available at `http://aspics.krzakala.org`.

The BP approximation lies on the independence of the entries of the sensing matrix, a condition that is not generally fulfilled in the matrices considered in this section. In

Accepted Manuscript

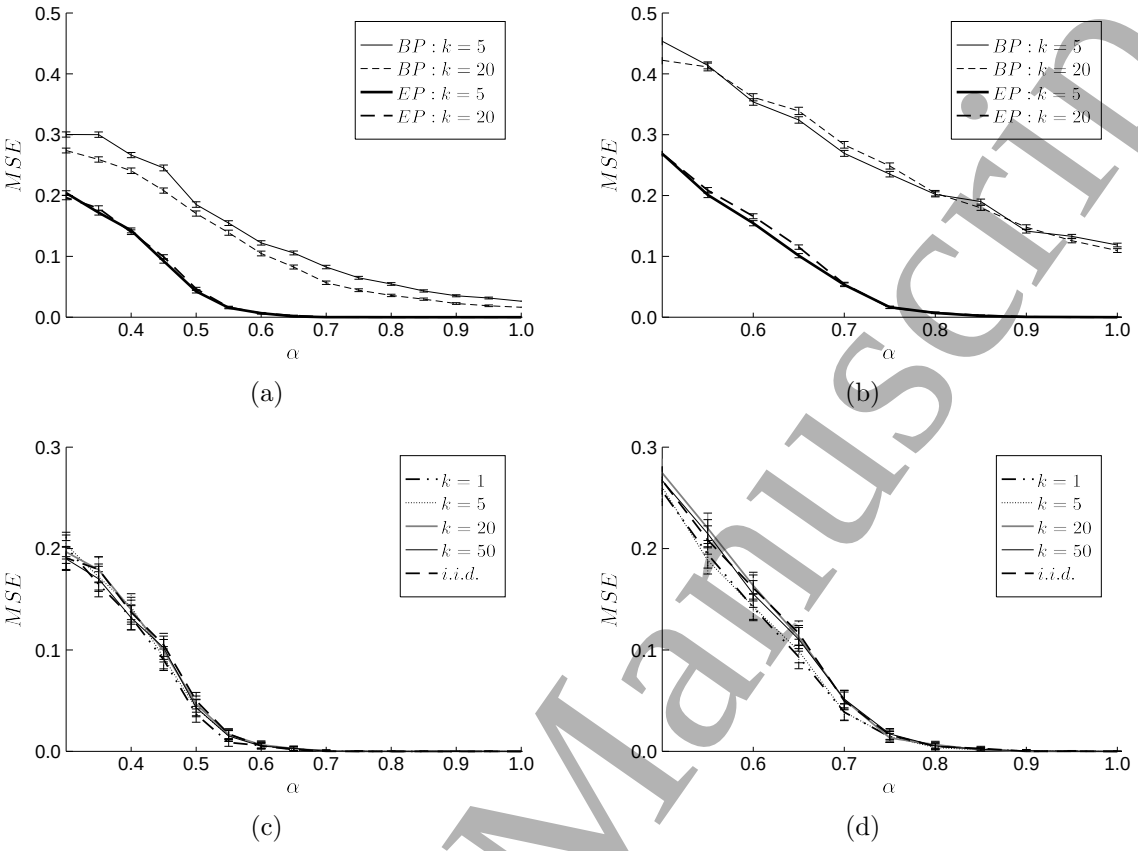


Figure 5. (a),(b) Comparison between the MSE obtained when reconstructing by means of EP and by means of EMBP in the case of correlated measurement matrices. The plot shows the MSE as evaluated over $N_t = 1000$ converged trials and the uncertainty has been estimated as $\sigma/\sqrt{N_t}$, where σ is the sample standard deviation. Regardless of the value of α, the EMBP algorithm is not able to reconstruct correctly the signal, whereas EP achieves zero MSE beyond a critical value $\alpha_c(\rho)$. (c),(d) MSE resulting from the reconstruction from both i.i.d. and correlated measurement matrices. Each point was evaluated over $N_t = 1000$ trials. Lower values of k correspond to more correlated measurements. The number of variables is $N = 50$ and the density of the signal is (a),(c) $\rho = 0.3$ and (b),(d) $\rho = 0.5$.

particular, for small values of k the covariances and the variances of $\mathbf{S}$ are of the same order of magnitude. However, as k increases, these variances become dominant with respect to the off-diagonal entries of $\mathbf{S}$ and the associated multivariate Gaussian measure becomes more and more similar to the product of independent univariate Gaussian distributions. In figures 5(a) and 5(b), we compare the MSE associated with EMBP and with EP when using correlated matrices and for $\rho = 0.3$, $\rho = 0.5$ respectively. The signal density is learned in both cases, using Expectation Maximization in the case of EMBP and using the free energy-based variational method described in Appendix C in the case of EP. Only converged trials have been taken into account. While BP fails to correctly reconstruct the signal and to infer the signal density in the presence of the correlations we introduced in the measurements, the performance of the EP reconstruction is unaffected. As we expected, EMBP performances improve as k increases. However, we note that at

Accepted Manuscript

low enough values of k, such as those we have considered, EMBP never achieves zero MSE, not even when $\alpha = 1$ (that is when we have as many equations as variables).

The fraction of converged trials is generally far lower in the case of EMBP than in the case of EP and decreases as N is increased. For example, in the case of $k = 1$, for $N = 50$, it is of the order of one in a thousand of simulated trials, two orders of magnitude lower than the fraction of converged EP simulations at the same values of N and k (not shown).

We also plot the MSE resulting from the EP reconstruction for i.i.d. measurement matrices and correlated measurement matrices that are constructed using equations (20), (21) and (22) for various values of k. By considering $k = 50$, $k = 20$, $k = 10$, $k = 5$ and $k = 1$, we obtain that the associated mean squared errors do not exhibit significant discrepancies, as shown in figures 5(c) and 5(d). This confirms that the EP inference of the signal is not altered with respect to the case of i.i.d. sensing matrices and retains its correct-incorrect reconstruction threshold.

Finally, we tested several algorithms for sparse reconstruction on linear systems with the same type of correlated measurement matrices considered so far. More precisely, these algorithms are Basis Pursuit [36], Orthogonal Matching Pursuit (OMP) [37], Regularized Orthogonal Matching Pursuit [38], Compressive Sampling Matching Pursuit (CoSaMP) [39], Subspace Pursuit [40], Smoothed L0 (SL0) [41], Approximate Message Passing (AMP) [13] and, again, Expectation Maximization Belief Propagation [24]. These algorithms are implemented in the C++ library KL1p [42]. This specific implementation makes use of the linear algebra library Armadillo [43, 44].

In order to compare the performance of these algorithms, we generated $N_t = 100$ random gaussian i.i.d. signals of length $N = 100$ and as many random correlated sensing matrices, with $k = 5$. For any given pair of signal $\boldsymbol{w}$ and measurement matrix $\mathbf{F}$, we attempted to recover the original signal by means of EP and of the algorithms included in KL1p. The results are presented in figure 6. As we can see in figure 6(a) and as further highlighted in the semi-logarithmic plot in figure 6(b), EP is the only algorithm exhibiting an incorrect-correct reconstruction phase transition, whereas all the other methods that we considered fail to retrieve the signal regardless of the value of α. In terms of running time, EP appears to be comparable to most of the other reconstruction techniques, as shown in figure 6(c).

4. Conclusions

We have proposed an EP-based scheme for efficient CS reconstruction whose computational complexity is dominated by a matrix inversion per iteration which requires $O(N^3)$ operations

By analyzing the reconstruction achieved by EP in the case of undersampled linear systems with random i.i.d. measurement matrices, we showed that EP exhibits a phase transition, in analogy to other message passing inspired algorithms. We numerically computed the threshold and the related phase transition diagram and found that signal

Accepted Manuscript

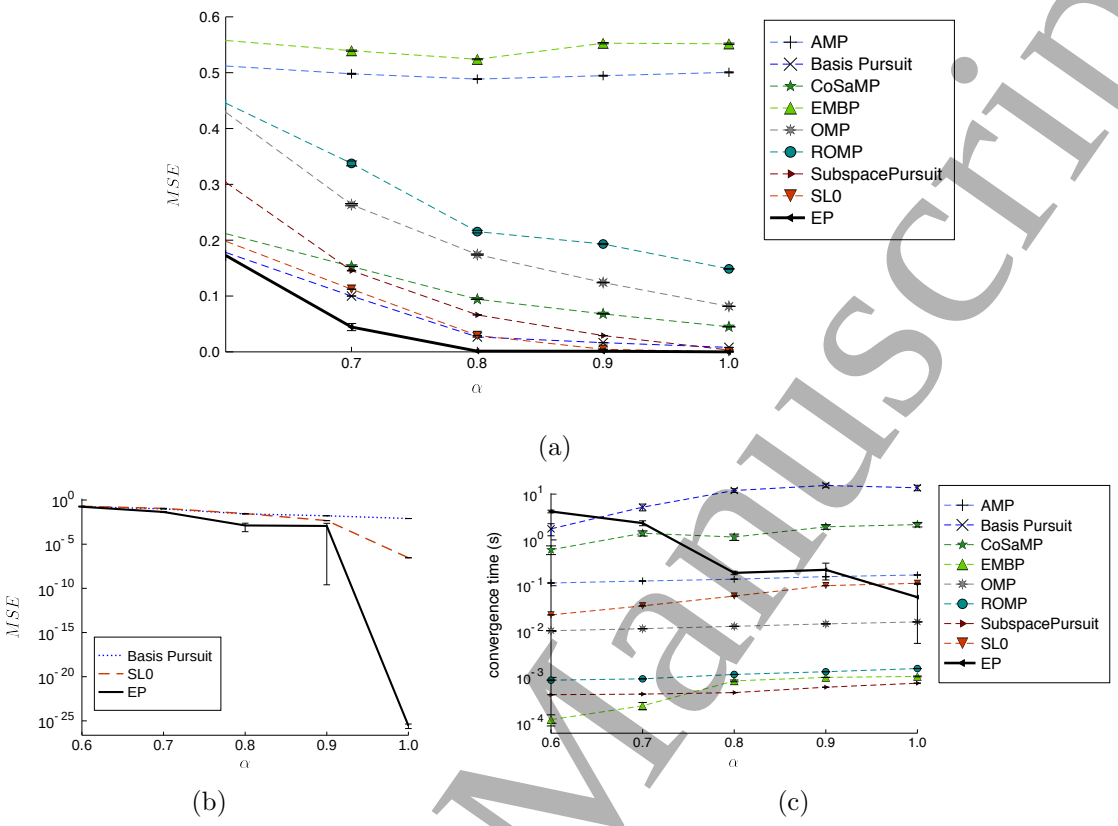


Figure 6. (a) Comparison of the MSE of various CS reconstruction algorithms in the presence of correlated measurement matrices with $k = 5$. The signals to be retrieved have density $\rho = 0.5$. (b) Semi-logarithmic plot of the reconstruction error of Basis Pursuit, SL0 and EP. (c) Comparison of the elapsed running time of the same reconstruction algorithms in the presence of correlated measurement matrices. In both figures, the parameters of the generated signals are given by $N = 100$, $\rho = 0.5$ and κ=5 and the total number of trials is $N_t = 100$.

reconstruction is possible below the L_1 minimization line (see figures 1, and 2).

Finally, we investigated the case of correlated measurement matrices and found that the EP threshold persists, implying that EP is still capable of accurately retrieving the signal beyond a critical α and that, contrary to the case of other reconstruction algorithms, it is robust against the presence of statistical structure in the measurements.

5. Acknowledgements

We acknowledge the kind hospitality of the Physics Department of the University of Havana where part of this work has been completed. This work has received support by INFERNET – Marie Skłodowska-Curie grant agreement number 734439 – and by the SmartData@PoliTO interdepartmental center of Politecnico di Torino.

Accepted Manuscript

Appendix A. Expectation Propagation in the zero temperature limit

In section 2, we implemented the linear constraints in equation (1) as the $\beta \to \infty$ limit of the multivariate Gaussian measure. We are going to show here how one can compute analytically this limit. Let us rewrite the previous formalism in a slightly different matrix notation. We will assume that the rank of the measurement matrix is maximum and equal to M, with $M < N$ (if this is not the case, we can easily remove the linearly dependent rows from the measurement matrix $\mathbf{F}$). Through Gaussian elimination, we can transform the matrix $\mathbf{F}$ to a row echelon form:

$$\mathbf{F}' = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{G} \end{array} \right] = \begin{bmatrix} 1 & & & G_{1,1} & \dots & G_{1,N-M} \\ & 1 & & G_{2,1} & \dots & G_{2,N-M} \\ & & \dots & \dots & \dots & \dots \\ & & 1 & G_{M,1} & \dots & G_{M,N-M} \end{bmatrix}$$

The structure of the linear constraint induced by the row echelon representation suggests to split the $\boldsymbol{w}$ variable into two sets of variables: the first M variables (dependent) and a second set of $N - M$ variables (independent). To do so, we define:

$$\boldsymbol{w} =: (\boldsymbol{w}^{(d)}, \boldsymbol{w}^{(i)}) = w_1^{(d)}, \dots, w_M^{(d)}, w_1^{(i)}, \dots, w_{N-M}^{(i)}$$

where $\boldsymbol{w}^{(d)} \in \mathbb{R}^M$, and $\boldsymbol{w}^{(i)} \in \mathbb{R}^{N-M}$. The linear constraint in equation (1) now reads:

$$\boldsymbol{w}^{(d)} + \mathbf{G}\boldsymbol{w}^{(i)} = \boldsymbol{y}'$$

where $\boldsymbol{y}' \in \mathbb{R}^M$ is the transformed measurement vector. Assuming, as in the previous subsection, a Gaussian prior for the $\boldsymbol{w}^{(i)}$ variables, there follows a Gaussian statistics on the $\boldsymbol{w}^{(d)}$ variables with consistent moments:

$$\bar{\boldsymbol{w}}^{(d)} = -\mathbf{G}\bar{\boldsymbol{w}}^{(i)} + \boldsymbol{y}'$$
$$\boldsymbol{\Sigma}_{\boldsymbol{w}^{(i)}} = \left(\mathbf{D}_{\boldsymbol{w}^{(i)}} + \mathbf{G}^T \mathbf{D}_{\boldsymbol{w}^{(d)}} \mathbf{G}\right)^{-1} \quad , \quad \boldsymbol{\Sigma}_{\boldsymbol{w}^{(d)}} = \mathbf{G}\boldsymbol{\Sigma}_{\boldsymbol{w}^{(i)}}\mathbf{G}^T$$

where the $\boldsymbol{D}_{\boldsymbol{w}^{(d)}}, \boldsymbol{D}_{\boldsymbol{w}^{(i)}}$ are the diagonal matrices whose entries are the inverses of the variances of the Gaussian priors associated with the dependent and independent variables, respectively. We note that from the previous relations the moments of the dependent $\boldsymbol{w}^{(d)}$ variables can be directly related to those of the $\boldsymbol{w}^{(i)}$ variables, which allows us to compute everything in terms of the inverse of a smaller matrix of size $(N - M) \times (N - M)$ compared to the finite temperature case.

At this point, the parameters $\boldsymbol{a}$ and $\boldsymbol{d}$ can be updated by moment matching as described in section 2.

Appendix B. Moments of the spike-and-slab prior

We use the spike-and-slab prior, defined as:

$$\psi(w_n) = (1 - \rho)\delta(w_n) + \frac{\rho}{\sqrt{2\pi\lambda}} e^{-\frac{w_n^2}{2\lambda}}, \tag{B.1}$$

Accepted Manuscript

where $\rho = K/N$ is the density of the signal $\boldsymbol{w}$. For spike-and-slab priors, the n-th marginal of the tilted distribution (9) is given by:

$$Q^{(n)}(w_n) = \frac{1}{Z_{Q^{(n)}}} \tilde{Q}_n(w_n)\psi_n(w_n) \tag{B.2}$$

in which:

$$\tilde{Q}_n(w_n) = \frac{1}{\sqrt{2\pi\Sigma'_n}} e^{-\frac{(w_n-\bar{w}'_n)^2}{2\Sigma'_n}}, \tag{B.3}$$

where we have denoted $\left(\bar{w}_{(n)}\right)_n$ and $\left(\Sigma_{(n)}\right)_{n,n}$ by $\bar{w}'_n$ and Σ'_n, respectively, in order to simplify the notation.

The partition function in (B.2) reads:

$$Z_{Q^{(n)}} = (1-\rho)\frac{1}{\sqrt{2\pi\Sigma'_n}} e^{-\frac{\bar{w}'^2_n}{2\Sigma'_n}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma'_n)}} e^{-\frac{1}{2}\frac{\bar{w}'^2_n}{\lambda+\Sigma'_n}}, \tag{B.4}$$

and the first and second moment of w_n with respect to $Q^{(n)}$ are given by:

$$\langle w_n \rangle_{Q^{(n)}} = \frac{1}{Z_{Q^{(n)}}} \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma'_{(n)})}} \frac{\lambda\bar{w}'_n}{\lambda+\Sigma'_n} e^{-\frac{1}{2}\frac{\bar{w}'^2_n}{\lambda+\Sigma'_n}}, \tag{B.5}$$

and by:

$$\langle w_n^2 \rangle_{Q^{(n)}} = \frac{1}{Z_{Q^{(n)}}} \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma'_n)}} \left(\frac{\lambda\Sigma'_n(\lambda+\Sigma'_n) + \lambda^2\bar{w}'^2_n}{(\lambda+\Sigma'_n)^2} \right) e^{-\frac{1}{2}\frac{\bar{w}'^2_n}{\lambda+\Sigma'_n}}, \tag{B.6}$$

respectively.

Appendix C. Learning of the density parameter of the prior

Appendix C.1. The EP Free Energy function

Let us consider equation (9) and rewrite it through equation (6) as:

$$\begin{aligned} Q^{(n)}(\boldsymbol{w}|\mathbf{F},\boldsymbol{y}) &:= \frac{1}{Z_{Q^{(n)}}} e^{-\beta\frac{(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})^T\cdot(\boldsymbol{y}-\mathbf{F}\boldsymbol{w})}{2}} \psi_n(w_n) \prod_{l\neq n} \phi_l(w_l; a_l, d_l) \\ &= \frac{Z_Q}{Z_{Q^{(n)}}} Q(\boldsymbol{w}|\mathbf{F},\boldsymbol{y}) \frac{\psi_n(w_n)}{\phi_n(w_n)}\,. \end{aligned} \tag{C.1}$$

We define:

$$Z_{EP} = Z_Q \prod_{n=1}^{N} \frac{Z_{Q^{(n)}}}{Z_Q} = \frac{\prod_{n=1}^{N} Z_{Q^{(n)}}}{Z_Q^{N-1}}, \tag{C.2}$$

from which the so-called *EP free energy* follows:

$$F_{EP} = (N-1)\log Z_Q - \sum_{n=1}^{N} \log Z_{Q^{(n)}}. \tag{C.3}$$

Accepted Manuscript

The converged means $\boldsymbol{a}$ and variances $\boldsymbol{d}$ of EP, which fulfill the moment matching conditions (11) for all $n = 1, \ldots, N$, are fixed points of the EP free energy, where the latter are obtained from:

$$0 = \frac{\partial F_{EP}}{\partial a_n} = (N-1)\langle w_n \rangle_Q - \sum_{l \neq n} \langle w_n \rangle_{Q^{(l)}}, \tag{C.4}$$

$$0 = \frac{\partial F_{EP}}{\partial d_n} = (N-1)\langle w_n^2 \rangle_Q - \sum_{l \neq n} \langle w_n^2 \rangle_{Q^{(l)}}, \tag{C.5}$$

for $n = 1, \ldots, N$. In order to show this, we shall prove that the moment matching conditions imply that (C.4) and (C.5) are satisfied. We have for $\langle w_n \rangle_{Q^{(l)}}$ and $\langle w_n^2 \rangle_{Q^{(l)}}$:

$$\begin{aligned} \langle w_n^\alpha \rangle_{Q^{(l)}} &= \int \frac{Z_Q}{Z_{Q^{(l)}}} Q(\boldsymbol{w}) \frac{\psi_l(w_l)}{\phi_l(w_l)} w_n^\alpha d\boldsymbol{w} = \int Q(\boldsymbol{w}) \frac{Q^{(l)}(w_l)}{Q(w_l)} w_n^\alpha d\boldsymbol{w} = \\ &= \left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha dw_n \right\rangle_{Q^{(l)}(w_l)}, \quad \alpha = 1, 2 \end{aligned}$$

In the last equality, for $\alpha = 1$ ($\alpha = 2$), the integral being averaged w.r.t. $Q^{(l)}(w_l)$ is the first (second) moment of w_n, conditioned on w_l and computed w.r.t. Q. These moments depend on w_l through the mean (squared mean) of $Q(w_n|w_l)$. As such mean depends linearly on w_l, $\langle w_n \rangle_{Q(w_n|w_l)}$ and $\langle w_n^2 \rangle_{Q(w_n|w_l)}$ depend on w_l linearly and quadratically, respectively. Therefore, for $\alpha = 1, 2$, by the moment matching conditions, we have that:

$$\left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha dw_n \right\rangle_{Q^{(l)}(w_l)} = \left\langle \int_{-\infty}^{+\infty} \frac{Q(w_n, w_l)}{Q(w_l)} w_n^\alpha dw_n \right\rangle_{Q(w_l)}, \tag{C.6}$$

implying that $\langle w_n^\alpha \rangle_{Q^{(l)}} = \langle w_n^\alpha \rangle_Q$ and thus that the conditions (C.4) and (C.5) are identically fulfilled.

Appendix C.2. Learning of the density

We are interested in finding the maximum likelihood value of the density parameter ρ which appears in the prior factors $\psi_n(w_n)$. The likelihood of the parameters of the prior is given by:

$$P(\boldsymbol{y}|\rho, \lambda) = \int d\boldsymbol{w} P(\boldsymbol{y}, \boldsymbol{w}|\rho, \lambda) = \tag{C.7}$$

$$= \int d\boldsymbol{w} P(\boldsymbol{y}|\boldsymbol{w}) P(\boldsymbol{w}|\rho, \lambda) = Z(\rho, \lambda) \tag{C.8}$$

and maximizing this likelihood corresponds to minimizing the associated free energy $F(\rho, \lambda) = -\log Z(\rho, \lambda)$.

At the fixed point of EP, the free energy is approximated by F_{EP} and the parameters can be learned by gradient descent. In particular, we have for the signal density ρ:

$$\rho^{(t+1)} \leftarrow \rho^{(t)} - \eta \frac{\partial F_{EP}}{\partial \rho}, \tag{C.9}$$

where t denotes the current iteration and η is a learning rate. In the simulations of this paper, we have taken $\eta = 5 \times 10^{-4}$.

Accepted Manuscript

The parameters of the prior enter in the EP free energy through the contributions associated with each of the tilted distributions. Such contributions read:

$$F_{Q^{(n)}} = -\log Z_{Q^{(n)}} = -\log\left(\int \tilde{Q}^{(n)}(\boldsymbol{w}|\boldsymbol{y})\psi_n(w_n)d\boldsymbol{w}\right) \tag{C.10}$$

for

$$\tilde{Q}^{(n)}(\boldsymbol{w}|\boldsymbol{y}) = e^{-\frac{1}{2}(\boldsymbol{w}-\bar{\boldsymbol{w}}_{(n)})^T\Sigma_{(n)}^{-1}(\boldsymbol{w}-\bar{\boldsymbol{w}}_{(n)})}. \tag{C.11}$$

Therefore, we have for $\frac{\partial F_{EP}}{\partial \rho}$:

$$\frac{\partial F_{EP}}{\partial \rho} = \sum_{n=1}^{N} \frac{\partial F_{Q^{(n)}}}{\partial \rho}, \tag{C.12}$$

where:

$$\frac{\partial F_{Q^{(n)}}}{\partial \rho} = -\frac{1}{Z_{Q^{(n)}}} \int \tilde{Q}_n(w_n) \frac{\partial \psi_n}{\partial \rho}(w_n) dw_n, \tag{C.13}$$

and:

$$\frac{\partial \psi_n}{\partial \rho}(w_n) = -\delta(w_n) + \frac{1}{\sqrt{2\pi\lambda}} e^{-\frac{w_n^2}{2\lambda}}, \tag{C.14}$$

yielding:

$$\frac{\partial F_{EP}}{\partial \rho} = \sum_{n=1}^{N} \frac{\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} - \frac{1}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}}{(1-\rho)\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}}. \tag{C.15}$$

By taking the derivative w.r.t. ρ one more time, we see that F_{EP} is a strictly convex function of ρ for $\lambda > 0$:

$$\frac{\partial^2 F_{EP}}{\partial \rho^2} = \sum_{n=1}^{N} \left[\frac{\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} - \frac{1}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}}{(1-\rho)\frac{1}{\sqrt{2\pi\Sigma_{n,n}}} e^{-\frac{\bar{w}_n^2}{2\Sigma_{n,n}}} + \frac{\rho}{\sqrt{2\pi(\lambda+\Sigma_{n,n})}} e^{-\frac{1}{2}\frac{\bar{w}_n^2}{\lambda+\Sigma_{n,n}}}} \right]^2, \tag{C.16}$$

which guarantees that the sought value of ρ is unique at fixed $\bar{w}_n$ and $\Sigma_{n,n}$.

References

[1] E. J. Candès and J. K. Romberg. "Signal recovery from random projections". In: *Computational Imaging III*. Vol. 5674. International Society for Optics and Photonics. 2005, pp. 76–87.

[2] D. L. Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[3] E. J. Candès and Michael B Wakin. "An introduction to compressive sampling [a sensing/sampling paradigm that goes against the common knowledge in data acquisition]". In: *IEEE signal processing magazine* 25.2 (2008), pp. 21–30.

Accepted Manuscript

[4] Z. Zhang et al. "A survey of sparse representation: algorithms and applications". In: *IEEE access* 3 (2015), pp. 490–530.

[5] M. Mishali et al. "Xampling: Analog to digital at sub-Nyquist rates". In: *IET circuits, devices & systems* 5.1 (2011), pp. 8–20.

[6] M. Mézard, G. Parisi, and Zecchina. R. "Analytic and Algorithmic Solution of Random Satisfiability Problems". In: *Science* 297.5582 (2002), pp. 812–815.

[7] R. Mulet et al. "Coloring random graphs". In: *Physical review letters* 89.26 (2002), p. 268701.

[8] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

[9] Y. Kabashima, T. Wadayama, and T. Tanaka. "A typical reconstruction limit for compressed sensing based on L p -norm minimization". In: *Journal of Statistical Mechanics: Theory and Experiment* 2009.09 (2009), p. L09003.

[10] M. Mézard, G. Parisi, and M. Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*. Vol. 9. World Scientific Publishing Company, 1987.

[11] E. J. Candès and T. Tao. "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

[12] E. J. Candès. "The restricted isometry property and its implications for compressed sensing". In: *Comptes Rendus Mathematique* 346.9 (2008), pp. 589 –592.

[13] D. L. Donoho, A. Maleki, and A. Montanari. "Message-passing algorithms for compressed sensing". In: *Proceedings of the National Academy of Sciences* 106.45 (2009), pp. 18914–18919.

[14] T. L. N. Nguyen and Y. Shin. "Deterministic Sensing Matrices in Compressive Sensing: A Survey". In: *The Scientific World Journal* 2013 (2013).

[15] L. Applebaum et al. "Chirp sensing codes: Deterministic compressed sensing measurements for fast recovery". In: *Applied and Computational Harmonic Analysis* 26.2 (2009), pp. 283 –290.

[16] K. Ni et al. "Image reconstruction by deterministic compressed sensing with chirp matrices". In: vol. 7497. 2009.

[17] S. D. Howard, A. R. Calderbank, and S. J. Searle. "A fast reconstruction algorithm for deterministic compressive sensing using second order reed-muller codes". In: *2008 42nd Annual Conference on Information Sciences and Systems*. 2008, pp. 11–15.

[18] E. J. Candès and T. Tao. "Near-Optimal Signal Recovery From Random Projections: Universal Encoding Strategies?" In: *IEEE Transactions on Information Theory* 52.12 (2006), pp. 5406–5425.

Accepted Manuscript

[19] E. J. Candès, J. Romberg, and T. Tao. "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509.

[20] P. Margosian, G. DeMeester, and H. Liu. "Partial Fourier Acquisition in MRI". In: Mar. 2007.

[21] C. Ravazzi, G. Coluccia, and E. Magli. "Image reconstruction from partial Fourier measurements via curl constrained sparse gradient estimation". In: Mar. 2017, pp. 4745–4749.

[22] L. Anitori et al. "Design and Analysis of Compressed Sensing Radar Detectors". In: *IEEE Transactions on Signal Processing* 61 (Feb. 2013), pp. 813–827.

[23] R. Holger. "Circulant and Toeplitz matrices in compressed sensing". In: *CoRR* abs/0902.4394 (2009).

[24] F. Krzakala et al. "Statistical-Physics-Based Reconstruction in Compressed Sensing". In: *Phys. Rev. X* 2 (2 2012), p. 021005.

[25] F. Krzakala et al. "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.08 (2012), P08009.

[26] T. J. Mitchell and J. J. Beauchamp. "Bayesian Variable Selection in Linear Regression". In: *Journal of the American Statistical Association* 83.404 (1988), pp. 1023–1032.

[27] M. Opper and O. Winther. "Gaussian processes for classification: mean-field algorithms". eng. In: *Neural Computation* 12.11 (2000), pp. 2655–2684.

[28] M. Opper and O. Winther. "Adaptive and self-averaging Thouless-Anderson-Palmer mean-field theory for probabilistic modeling". In: *Physical Review E* 64.5 (2001), p. 056131.

[29] T. P. Minka. "Expectation propagation for approximate Bayesian inference". In: *Proceedings of the Seventeenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 2001, pp. 362–369.

[30] A. Braunstein, A. P. Muntoni, and A. Pagnani. "An analytic approximation of the feasible space of metabolic networks". In: *Nature communications* 8 (2017), p. 14915.

[31] A. P. Muntoni et al. "Non-convex image reconstruction via Expectation Propagation". In: *arXiv preprint arXiv:1809.03958* (2018).

[32] Matthias W. Seeger. "Bayesian inference and optimal design for the sparse linear model". In: *The Journal of Machine Learning Research* 9 (2008), pp. 759–813.

[33] José Miguel Hernández-Lobato, Daniel Hernández-Lobato, and Alberto Suárez. "Expectation propagation in linear regression models with spike-and-slab priors". In: *Machine Learning* 99.3 (2015), pp. 437–487.

[34] A. Braunstein, G. Catania, and L. Dall'Asta. "Loop corrections in spin models through density consistency". In: *arXiv preprint arXiv:1810.10602* (2018).

Accepted Manuscript

[35] Y. Kabashima, T. Wadayama, and T. Tanaka. "Erratum: A typical reconstruction limit of compressed sensing based on L p -norm minimization". In: *Journal of Statistical Mechanics: Theory and Experiment* 2012.07 (2012), E07001.

[36] E. J. Candès and J. Romberg. "l1-magic: Recovery of sparse signals via convex programming". In: *URL: www.acm.caltech.edu/l1magic/downloads/l1magic.pdf* 4 (2005), p. 14.

[37] J. A. Tropp and A. C. Gilbert. "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit". In: *IEEE Transactions on Information Theory* 53.12 (2007), pp. 4655–4666.

[38] D. Needell and R. Vershynin. "Signal Recovery From Incomplete and Inaccurate Measurements Via Regularized Orthogonal Matching Pursuit". In: *IEEE Journal of Selected Topics in Signal Processing* 4.2 (2010), pp. 310–316.

[39] D. Needell and J. A. Tropp. "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples". In: *Applied and Computational Harmonic Analysis* 26.3 (2009), pp. 301 –321.

[40] W. Dai and O. Milenkovic. "Subspace Pursuit for Compressive Sensing Signal Reconstruction". In: *IEEE Transactions on Information Theory* 55.5 (2009), pp. 2230–2249.

[41] H. Mohimani, M. Babaie-Zadeh, and C. Jutten. "A Fast Approach for Overcomplete Sparse Decomposition Based on Smoothed ℓ^0 Norm". In: *IEEE Transactions on Signal Processing* 57.1 (2009), pp. 289–301.

[42] R. Gebel. *KL1p – A portable C++ library for Compressed Sensing*. 2012.

[43] C. Sanderson and R. Curtin. "A User-Friendly Hybrid Sparse Matrix Class in C++". In: *Lecture Notes in Computer Science (LNCS)* 10931 (2018), pp. 422–430.

[44] C. Sanderson and R. Curtin. "Armadillo: a template-based C++ library for linear algebra". In: *Journal of Open Source Software* 1 (2016), p. 26.

Accepted Manuscript