Analysis of Bayesian Compressive Sensing

Navaneetha Krishna Madan Gopal

Edward S. Rogers Sr. Department of Electrical Computer Engineering
University of Toronto
Toronto, Ontario, Canada
Email: navaneeth.madan@mail.utoronto.ca

***Abstract*—Shannon-Nyquist sampling theorem states that a signal can be reconstructed without error when it is sampled at twice the bandwidth or twice the maximum frequency of the signal. By way of compressive sensing, a sparse signal can be reconstructed using very few samples which is far less than the sampling rate required by the Shannon-Nyquist theorem. There are many approaches in compressive sensing like iterative relaxation, greedy algorithms and Bayesian methods. Here an overview of Bayesian compressive sensing is adopted, and its performance compared with other approaches.**

***Index Terms*—Compressed Sensing, Relevance Vector Machine, Bayesian Compressive Sensing**

I. Introduction

In the present day, ever more precise data acquisition systems are providing large amounts of data the storage of which has become a major challenge. As a result, various compression methods are employed to compress the data into more manageable sizes to facilitate storage and fast retrieval. This is the sample and compress philosophy. On the other hand, compressed sensing is based on the philosophy of compress while sampling. This method acquires few samples at the input and provides adequate fidelity reconstruction of the data at the output. The process of sampling is represented by the equation [1]

$$\mathbf{y} = \phi\mathbf{x} + \mathbf{n} \tag{1}$$

Here $\mathbf{x}$ is a sparse or compressible signal of length N. In a sparse K sparse signal, there are only K non-zero values while a compressible signal has exponentially decaying values, where the largest K values are considered with $K << N$. $\mathbf{n}$ is a length K noise vector. y is the length K vector of compressed samples. ϕ is an KxN matrix used to obtain the samples. This is the process of encoding the sparse vector. The operating equation for reconstruction of the sparse vector is [1]

$$min \parallel x \parallel_1 \; subject \; to \parallel y - \phi x \parallel_2 \leq \epsilon \tag{2}$$

The l2 norm is not considered as it does not encode sparsity rather it encodes the power in the signal. The l0 norm (number of non-zero values in the vector) is the ideal constraint but solving the optimization problem with such a constraint turns out to be an NP problem. Hence, a sub optimal solution which is the l1 norm is considered, and the problem is converted into a convex optimization problem. Obtaining the optimal solution to such a problem has been investigated thoroughly over the last few decades with linear programming approaches. Since the sensing matrix is not full rank, there are infinite

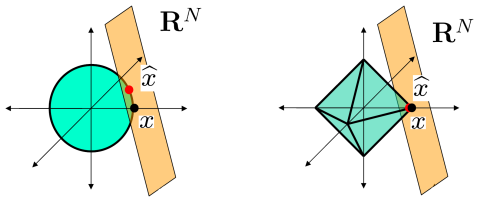

Fig. 1. Solutions using $L2$ and $L1$ norms [2]

values of x which when multiplied with it, give the same y. In N-dimensional space, these solutions form a hyper-plane oriented at a random angle determined by the sensing matrix. The sparse solutions to this exist near the coordinate axes. Geometrically, an L2 norm approach is equivalent to inflating an l2 ball from the origin until it touches the hyper-plane. Since it is spherical, it almost always does not touch the hyper-plane at a point close to the axes. This shows that the l2 norm approach is not optimal to encode sparsity. The l0 norm geometrically is the points of intersection of the hyper-plane with the N axes. Though this is the ideal solution, finding these points of intersection is an NP problem. Hence the l1 norm approach is adopted. As the l1 ball is pointed along the axes, it provides the sparse vectors as solutions. But for these methods to be successful, the sensing matrix must satisfy the Restricted Isometry Property (RIP) [1].

$$(1-\delta) \parallel x \parallel_2^2 \quad \leq \parallel \phi x \parallel_2^2 \leq (1+\delta) \parallel x \parallel_2^2 \tag{3}$$

The resultant vector of the multiplication of a 2S-sparse vector with a matrix which satisfies this property, has a length which is approximately same as the original 2S-sparse vector. Finding matrices which satisfy this property is an NP complete complete. But, most random matrices are known to satisfy the RIP for sparse signals. Classical compressive sensing techniques require the sensing matrix to satisfy this property. Hence, they use a KxN random matrix as the sensing matrix. The sensing matrix is needed for the reconstruction of the sparse signal. Storing (and transmitting this to the receiver in the case of communications) this random matrix is a major drawback of such a system. Bayesian compressive sensing methods on the

other hand, do not require the matrix to adhere to the RIP. Hence, a correlated matrix can be used for both encoding and reconstruction of the sparse signal. The correlation between the rows of the sensing matrix can be exploited to reduce the storage space requirement of the sensing matrix. In the Bayesian compressive sensing framework, the fact that the input signal is sparse is used to define a prior on the data. A sparsity promoting prior is defined on the input signal and the goal is to estimate the parameters of this prior. As a result of this process, the full posterior density of the reconstructed signal is obtained unlike the case of the classical approaches which avail a point estimate of the signal.

II. Relevance Vector Machine [3]

This is one of the methods of implementing Bayesian sparse recovery. This method utilizes the fact that most of the values in a sparse vector are highly peaked around zero, meaning their value is very close to, if not zero. The other values are called the relevant values and this method identifies these using a Bayesian framework. This approach also provides full posterior distribution of the output unlike most other methods like basis pursuit, orthogonal basis pursuit, matching pursuit to name a few. The latter methods provide a point estimate while the Relevance Vector machine avails error bars at the output. This provides a sense of confidence in the output. Using these error bars, the accuracy and reliability of future measurements or calculations based on the output of this algorithm can also be gauged.

III. Problem Setup

For the simulations, the following equation is used to generate the sampled vector y. Here y is a Kx1 vector, ϕ is a KxN correlated sensing matrix, x is an Nx1 sparse vector containing S spikes having values of either +1 or -1. n is a Kx1 noise vector which is normally distributed with variance sigma2. Providing y as input, the goal is to generate x as the output.

IV. Bayesian Analysis [4]

The first step in the Bayesian process is defining a prior on the data. Since the input vector is sparse, a sparsity promoting prior like Laplace can be used [5]. But since Laplace and Gaussian distributions are not conjugate, calculating the posterior probability becomes intractable. As a result of this, a different prior is needed which can ease the computation.

A. Hierarchical prior

Since working with the Laplace prior in intractable, a different prior called hierarchical prior is used. Here, instead of defining a common prior for the entire input, each entry of the input vector is modelled as a Gaussian random variable with the i^{th} entry having a precision α_i. On running the relevance vector machine, the precisions of the relevant entries stabilizes at a finite value while the precisions of the other entries tend to infinity. This signifies the peaking at the zero. Each of these α_is is modelled as a gamma random variable. The hyperparameters the gamma distribution are set to be zero to achieve a uniform non-informative prior. This model is called a hierarchical prior as the distribution is modelled in layers. The reason for expanding the problem to include more hyperparameters is to make sure the problem becomes tractable.

$$p(\mathbf{y}|\mathbf{x},\alpha_0) \;=\; (2\pi\alpha_0^{-1})^{-N/2} exp\{\alpha_0 \parallel \mathbf{y} - \phi\mathbf{x} \parallel^2\} \tag{4}$$

$$p(\mathbf{x}|\boldsymbol{\alpha}) = \prod_{i=1}^{N} \mathcal{N}(x_i|0,\alpha_i) \tag{5}$$

The noise is also modelled as a Gamma random variable with the hyperparameters set to 0 to have a non-informative prior.

B. Bayesian Modelling

The posterior of the unknowns given the output can be written as:

$$p(\mathbf{x},\boldsymbol{\alpha},\alpha_0|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x},\boldsymbol{\alpha},\alpha_0)p(\mathbf{x},\boldsymbol{\alpha},\alpha_0)}{p(\mathbf{y})} \tag{6}$$

But, performing the integration to obtain the denominator is very complex analytically. Hence the posterior is expressed as

$$p(\mathbf{x},\boldsymbol{\alpha},\alpha_0) = p(\mathbf{x}|\mathbf{y},\boldsymbol{\alpha},\alpha_0)p(\boldsymbol{\alpha},\alpha_0|\mathbf{y}) \tag{7}$$

From [ref2], the second term is replaced by the delta function at its mode. The distribution of the first term is:

$$p(\mathbf{x}|\mathbf{y},\boldsymbol{\alpha},\alpha_0) = \frac{p(\mathbf{y}|\mathbf{x},\alpha_0)p(\mathbf{x}|\boldsymbol{\alpha})}{p(\mathbf{y}|\boldsymbol{\alpha},\alpha_0)} \tag{8}$$

On simplification,

$$p(\mathbf{x}|\mathbf{y},\boldsymbol{\alpha},\alpha_0) = (2\pi)^{-(N-1)/2}|\boldsymbol{\Sigma}|^{-1/2}\; exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{9}$$

where,

$$\boldsymbol{\Sigma} = (\alpha_0\boldsymbol{\phi}^T\boldsymbol{\phi} + \mathbf{A})^{-1} \tag{10}$$

$$\boldsymbol{mu} = \alpha_0\boldsymbol{\Sigma}\boldsymbol{\phi}^T\mathbf{y} \tag{11}$$

The goal of maximizing precisions of the sparse vector components and noise given the observation. Given uniform hyperparameters, this is achieved by maximizing:

$$p(\mathbf{y}|\boldsymbol{\alpha},\alpha_0) = \int p(\mathbf{y}|\mathbf{x},\alpha_0)p(\mathbf{x}|\boldsymbol{\alpha}) \tag{12}$$

$$p(\mathbf{y}|\boldsymbol{\alpha},\alpha_0) = (2\pi)^{-N/2}|\alpha_0^{-1}\mathbf{I}+\boldsymbol{\phi}\mathbf{A}^{-1}\boldsymbol{\phi}^T|^{-1/2}exp\{-\frac{1}{2}\mathbf{y}^T(\alpha_0^{-1}\mathbf{I}+\boldsymbol{\phi}\mathbf{A}^{-1} \tag{13}$$

The values of precisions and variance cannot be obtained in a closed form, hence an iterative algorithm is use. From [McKay reference], this is achieved by

$$\alpha_i^{new} = \frac{\gamma_i}{\mu_i^2} \tag{14}$$

Where γ is

$$\gamma_i = 1 - \alpha_i\Sigma_{ii} \tag{15}$$

The noise variance is got using the equation

$$(\alpha_0^{-1})^{new} = \frac{\parallel \mathbf{y} - \boldsymbol{\phi}\boldsymbol{\mu} \parallel^2}{N - \Sigma_i\gamma_i} \tag{16}$$

The parameter gamma is a measure of the well determinedness of a data point in that index position. If there is no signal, the value of this parameter tends to zero. In the absence of the signal, the corresponding diagonal element of the signal covariance matrix will have a very low value as a consequence of the high value of the precision. On multiplying this value by the large value of precision, the resultant will be close to 1. Thus, the value of gamma tends to zero.

C. Iterative Algorithm

Once the new values of the precisions and noise variance have been obtained, the signal covariance matrix and mean vector are recalculated, and the estimation process is continued. The algorithm to implement this method is given below,

Steps of the implemented algorithm:

1) Generate a sparse input vector $\mathbf{x}$
2) Generate the sensing matrix ϕ and introduce correlation using $\boldsymbol{\rho}$
3) Obtain the observation vector $\mathbf{y}$
4) Initialize the vector of signal precisions $\boldsymbol{\alpha}$ as a noninformative prior
5) Initialize the noise variance to a reasonable value
6) Obtain the signal covariance matrix and mean vector
7) Obtain new values of the signal precisions and noise variance
8) Repeat from step 6 for a predefined number of iterations

The output sparse vector is the signal mean vector and the variance at each of the signal values is given the estimated signal precision vector.

D. Optimization:

As the algorithm progresses, a large gradient in the values of the signal precisions is seen. This is reflected as a larger condition number of the matrix (y*phi*phiT + A) and as a result, the inverting of this matrix to obtain the signal covariance matrix becomes inaccurate and computationally intensive. [3] This process of optimization is used to overcome this issue. The suggested method of optimization involves thresholding the values of gamma after every iteration. The row(s) and column(s) of the signal covariance matrix corresponding to the positions where gamma falls below the threshold are removed. This also extends to removing the corresponding values from the mean vector and the corresponding columns of the sensing matrix. Using this method, convergence can be well defined. Convergence is said to have been achieved when there are no values of gamma that are below the threshold. The output of this algorithm will be a diminished signal mean vector and signal precisions vector. The final reconstructed output is obtained by plugging the values in the mean vector to their initial index values and making the remaining values of the mean vector zero. The variances are also mapped onto the corresponding positions in the final length N signal precision vector. Improvement: In implementing this method, it is observed that the value of N reduces after every iteration. Hence this new updated value of N is used to estimate the noise variance. Based on multiple MATLAB simulations it was observed that a better estimate with lesser variance was produced when the original value of N was considered. Hence, in the implementation, this modification has been done.

V. Simulations

The simulations have been carried out for a range of values of correlations introduced into the sensing matrix. The performance of this method is compared with two other methods of compressive sensing. The first one being basis pursuit implemented using l1 magic[give sue credit] and the second one is the Fast Relevance Vector Machine algorithm implemented by [give due credit]

The values of parameters chosen for simulations are:

- Length of sparse vector N = 512
- Sample size K = 100
- Number of impulses S = 10
- $\boldsymbol{\rho}$ = [0.99991 : 0.00001 : 0.99999] ^{T}

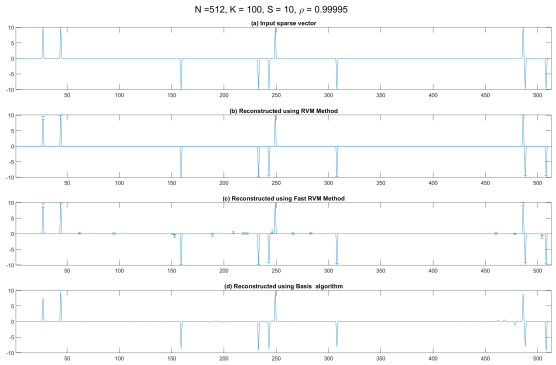

Fig. 2. Reconstructions at $\rho = 0.99995$

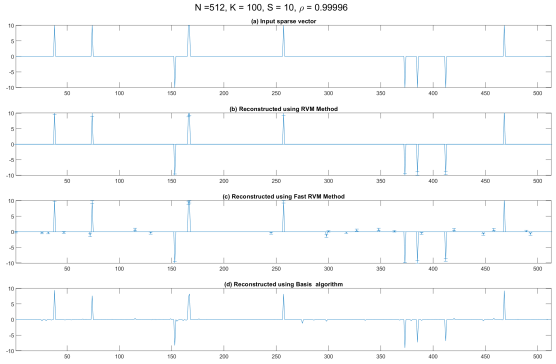

Fig. 3. Reconstructions at $\rho = 0.99996$

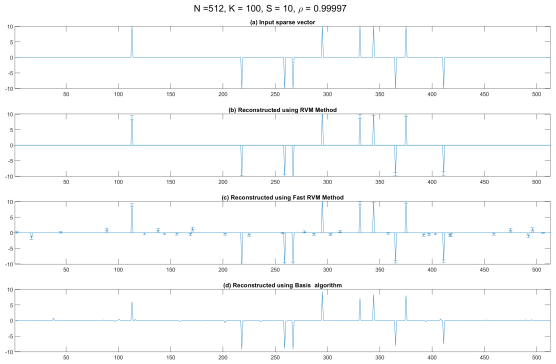

Fig. 4. Reconstructions at $\rho = 0.99997$

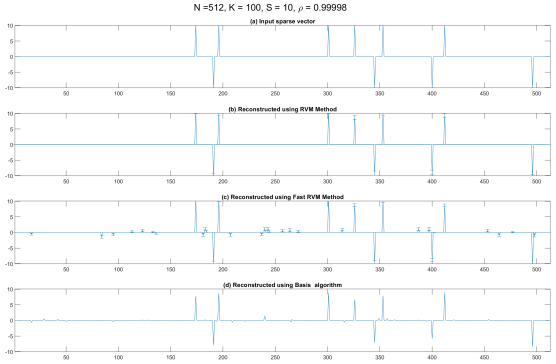

Fig. 5. Reconstructions at $\rho = 0.99998$

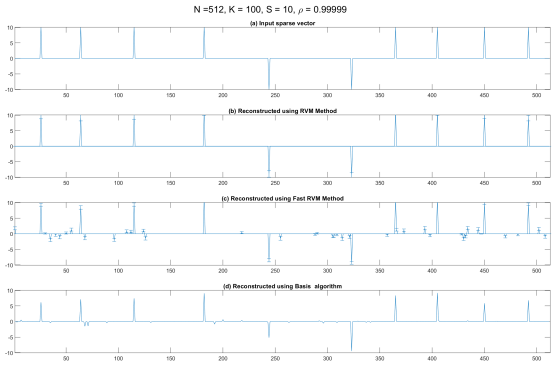

Fig. 6. Reconstructions at $\rho = 0.99999$

It is evident that the implemented algorithm is consistently performing better and providing high fidelity at even at high values of correlation. Three performance parameters for quantitative comparison [6]:

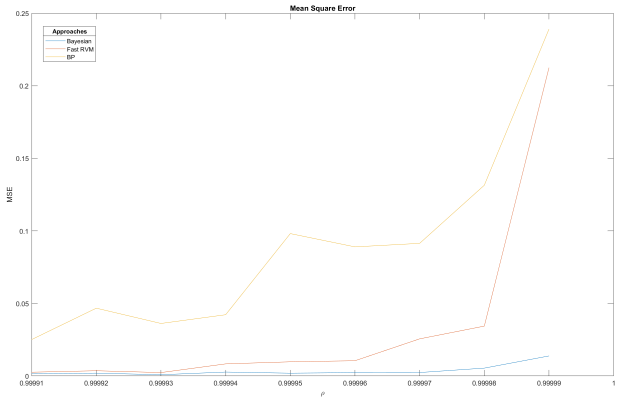

Fig. 7. Mean Square Error

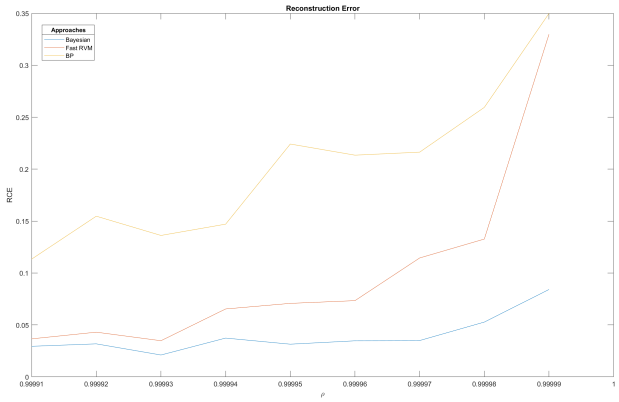

Fig. 8. Reconstruction Error

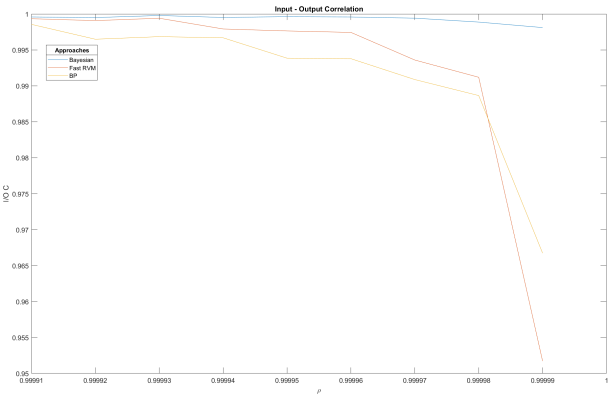

Fig. 9. Input Output Correlation

From the above plots, it is clear that for correlated sensing matrices, the basis pursuit algorithm has the poorest performance among the rest.

VI. Motivation and Acknowledgements

This has been submitted as a project for the course Detection and Estimation Theory (ECE 1521H F) under the guidance of Prof. R. Adve with the help if Ahmad Ali Khan.

References

[1] E. J. Candes, "The restricted isometry property and its implications for compressed sensing," *Applied and Computational Mathematics, California Institute of Technology*, 2008.

[2] R. Baraniuc, "Compressive Sensing," https://www.youtube.com/watch?v=zytez36XlCU, 2016.

[3] M. E. Tipping, "Sparse bayesian learning and the relevance vector machine," *Journal of Machine Learning Research 1*, 2001.

[4] Y. X. Shihao Ji and L. Carin, "Bayesian compressive sensing," Jan. 2007.

[5] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research 9*, 2008.

[6] H. E. G. Fatima Salahdine, Naima Kaabouch, "Bayesian compressive sensing with circulant matrix for spectrum sensing in cognitive radio networks," *Journal of Machine Learning Research 9*, 2016.