Bayesian Inference and Compressed Sensing

Solomon A. Tesfamicael and Faraz Barzideh

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/intechopen.70308

Abstract

This chapter provides the use of Bayesian inference in compressive sensing (CS), a method in signal processing. Among the recovery methods used in CS literature, the convex relaxation methods are reformulated again using the Bayesian framework and this method is applied in different CS applications such as magnetic resonance imaging (MRI), remote sensing, and wireless communication systems, specifically on multiple-input multiple-output (MIMO) systems. The robustness of Bayesian method in incorporating prior information like sparse and structure among the sparse entries is shown in this chapter.

Keywords: Bayesian inference, compressive sensing, sparse priors, clustered priors, convex relaxation

1. Introduction

In order to estimate parameters in a signal, one can apply wisdoms of the two schools of thoughts in statistics called the classical (also called the frequentist) and the Bayesian. These methods of computing are competitive with each other at times. The definition of probability is where the basic difference arises from. The frequentist define $P(A)$ as a long-run relative frequency with which A occurs in identical repeats of an experiment, whereas Bayesian defines $P(A|B)$ as a real number measure of the probability of a proposition A, given the truth of the information represented by proposition B. So under Bayesian theory, probability is considered as an extension of logic [1, 2]. Probabilities represent the investigator degree of belief—hence it is subjective. But this is not acceptable under classical theory, making it to be not flexible. To add on the differences, under the classical inference, parameters are not random, they are fixed and prior information is absent. But under the Bayesian, parameters are random variables, and prior information is an integral part, and the Bayesian has no excuse for that. Since one is free

to invent new estimators or confidence intervals or hypothesis test, adhockery exists and hence frequentist approach lack consistency whereas Bayesian theory is flexible and consistent [1–9]. Therefore, Bayesian inference is our main focus applied to a special paradigm in signal processing in this chapter.

After presenting the theoretical frameworks, Bayesian theory, CS, and convex relaxation methods in Section 2, the use of Bayesian inference in CS problem by considering two priors modeling the sparsity and clusteredness is shown in Section 3. In Section 4, we present three examples of applications that show the connection of the two theories, Bayesian and compressive sensing. In Section 5, the conclusion is given briefly.

2. Theoretical framework

2.1. Bayesian framework

For two random variables A and B, the product rule gives

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A) \tag{1}$$

and the famous Bayes' theorem provides

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}. \tag{2}$$

Using the same framework, consider model M_j and a vector of parameter $\boldsymbol{\theta}$. We infer what the model's parameter $\boldsymbol{\theta}$ might be, given the data, D, and a prior information I. Using Bayes' theorem, the probability of the parameters θ given model M_j, data D, and information I is given by

$$P(\boldsymbol{\theta}|D, M_j, I) = \frac{P(D|\boldsymbol{\theta}, M_j, I)P(\boldsymbol{\theta}|M_j, I)}{P(D|M_j, I)}, \tag{3}$$

where $P(\boldsymbol{\theta}|D, M_j, I)$, is posterior probability, $P(\boldsymbol{\theta}|M_j, I)$ is the non data information about $\boldsymbol{\theta}$, called prior probability distribution function, while $P(D|\boldsymbol{\theta}, M_j, I)$ is the density of the data conditional on the parameters of the model, called likelihood. $P(D|M_j, I)$ is called the evidence of model M_j, or the normalizing constant, given by:

$$P(D|M_j, I) = \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|M_j, I)P(D|\boldsymbol{\theta}, M_j, I)d\boldsymbol{\theta}. \tag{4}$$

$P(\boldsymbol{\theta}|D, M_j, I)$ is the fundamental interest for the first level of inference called model fitting. It is the task of inferring what the model parameters might be given the model and the data. Further, we can do inference on higher level, which is comparing models M_j. In the light of prior information I and data D, a given set of models $\{M_1, M_2, \cdots, M_n\}$ is most likely to be the

correct one. Now focusing on the first level of inference, we can ignore the normalizing constant in (3) since it has no relevance at this level of inference about the parameters θ. Hence we get:

$$P(\boldsymbol{\theta}|D, M_j, I) \propto P(D|\boldsymbol{\theta}, M_j, I)P(\theta|M_j, I). \tag{5}$$

The *posterior* probability is proportional to the *Prior* probability times the *Likelihood*. Eq. (5) is called Updating Rule [1, 3], in which the data allow us to update our prior views about θ, and as a result, we get the posterior which combines both the data and non-data information of θ. As an example for a binomial trial, let us have beta distribution as a prior and as a result we get posterior distribution which is beta distribution. **Figure 1** shows that the posterior density is taller and narrower than the prior density. It therefore favors strongly a smaller range of θ values, reflecting the fact that we now have more information. That is why inference based on the posterior distribution is superior to the one only based on the likelihood.

Now, we first find the maximum of the posterior distribution called maximum a posteriori (MAP). It defines the most probable value for the parameters denoted $\widehat{\theta}_{MP}$. MAP is related to Fisher's methods of maximum likelihood estimation (MLE), $\widehat{\theta}_{ML}$. If f is the sampling distribution of D, then the likelihood function of $D{:}\theta \mapsto f(D|\theta)$ and the maximum likelihood estimation of θ:

$$\boldsymbol{\theta}_{ML}(D) = \arg \max_{\boldsymbol{\theta}} f(D|\boldsymbol{\theta}). \tag{6}$$

But under Bayesian inference, let g be a prior distribution of $\boldsymbol{\theta}$, then the posterior distribution of $\boldsymbol{\theta}$ becomes

$$\boldsymbol{\theta}_{ML} \mapsto \frac{f(D|\boldsymbol{\theta})g(\boldsymbol{\theta})}{f(D)} \tag{7}$$

and the maximum a posteriori estimation of θ:

$$\widehat{\boldsymbol{\theta}}_{MP} = \frac{f(D|\boldsymbol{\theta})g(\boldsymbol{\theta})}{\int_{\boldsymbol{\vartheta}} f(D|\boldsymbol{\vartheta})g(\boldsymbol{\vartheta})d\boldsymbol{\vartheta}}$$

$$= \arg \max_{\boldsymbol{\theta}} f(D|\boldsymbol{\theta})g(\boldsymbol{\theta}) \tag{8}$$

Inference based on the posterior is not an easy task since it involves multiple integral, which are cumbersome to solve at times. However, it can be computed in several ways: Numerical optimization (like Conjugate gradient method, Newton method,…), modification of an expectation-maximization algorithm and others. As we can see it from (22) and (8), the difference between MLE and MAP is the prior distribution. The latter can be considered as a
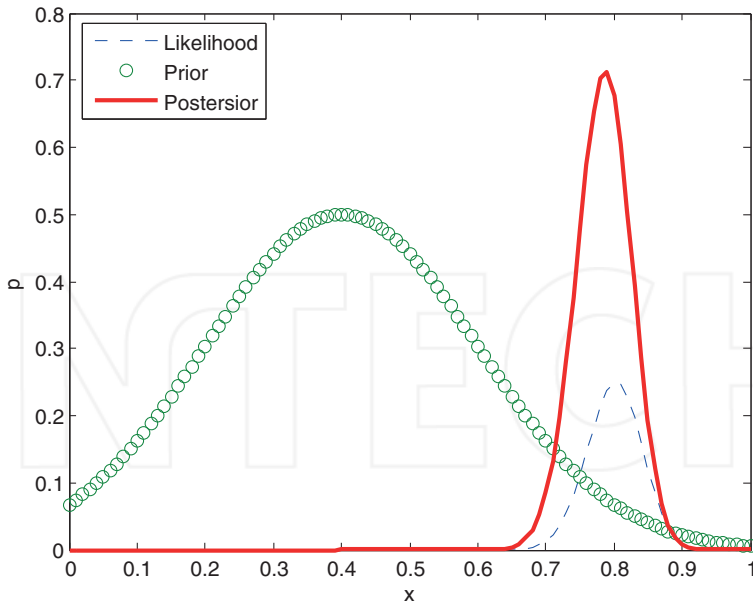
Figure 1. Figure showing the updating rule: the posterior synthesizes and compromises by favoring values between the maximum of the prior density and likelihood. The prior we had is challenged to shift by the arrival of little amount of data.

regularization of the former. Here we can summarize the posterior distribution by the value of the best fit parameters θ_{MP} and error bars (confidence intervals) on the best fit parameters. Error bars can be found from the curvatures of the posterior. To proceed further, we replace the random variable D and θ by vectors **y** and **x** and we assume prior distributions on **x** in the next section.

2.2. Compressive sensing

Compressive sensing (CS) is a paradigm to capture information at lower rate than the Nyquist-Shannon sampling rate when signals are sparse in some domain [10–13]. CS has recently gained a lot of attention due to its exploitation of signal sparsity. Sparsity, an inherent characteristic of many natural signals, enables the signal to be stored in a few samples and subsequently be recovered accurately.

As a signal processing scheme, CS follows a similar fashion: encoding, transmission/storing, and decoding. Focusing on the encoding and decoding of such a system with noisy measurement, the block diagram is given in **Figure 2**. At the encoding side, CS combines the sampling and compression stages of a traditional signal processing into one step by measuring few samples that contain maximum information about the signal. This measurement/sampling is done by linear projections using random sensing transformations as shown in the landmark papers by the authors mentioned above. Having said this, let us define the CS problem formally as follows:
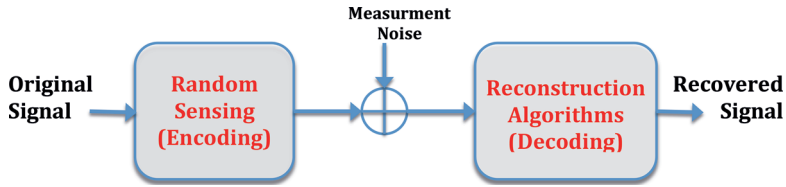
Figure 2. Blockdiagram for CS-based reconstruction.

Definition 1. (*The standard CS problem*)

Find the k-sparse signal vector $\mathbf{x} \in \mathbb{R}^N$ *provided the measurement vector* $\mathbf{y} \in \mathbb{R}^M$, *the measurement matrix* $\mathbf{A} \in \mathbb{R}^{M \times N}$ *and the under-determined set of linear equations as*

$$\mathbf{y} = \mathbf{Ax}, \tag{9}$$

where $k \ll M \ll N$.

One can ask again two of the questions here, in relation to the standard CS problem. First, how should we design the matrix $\mathbf{A}$ to ensure that it preserves the information in the signal $\mathbf{x}$? Second, how can we recover the original signal $\mathbf{x}$ from measurements $\mathbf{y}$ [14]? To address the first question, the solution for the CS problem presented here is dependent on the design of $\mathbf{A}$. This matrix can be considered as a transformation of the signal from the signal space to the measurement space, **Figure 3** [15]. There have been different criteria that matrix $\mathbf{A}$ should satisfy to have meaningful reconstruction. One of the main criteria is given in [11]. The authors defined the sufficient condition that matrix $\mathbf{A}$ should satisfy for the reconstruction of the signal $\mathbf{x}$. It is called the Restricted Isometric Property (RIP) and it is defined below.
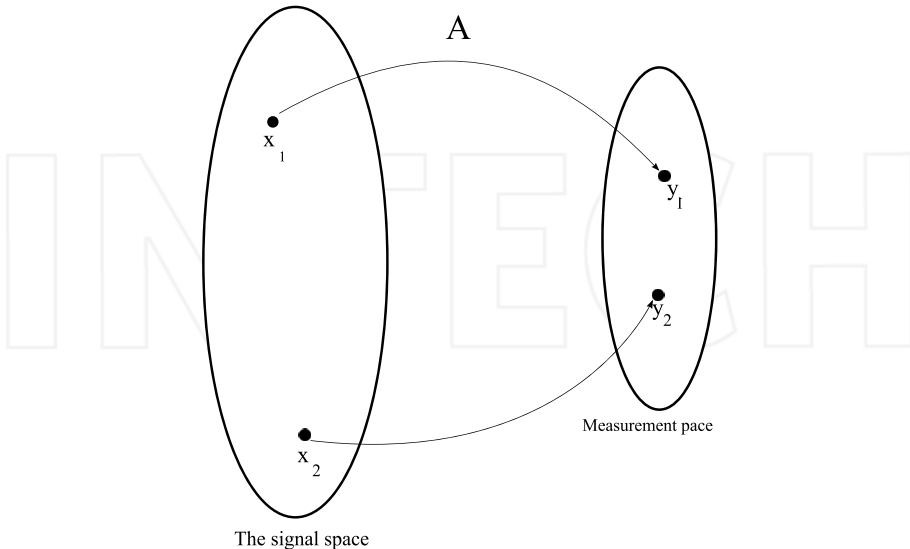

Figure 3. Transformation from the signal-space to the measurement-space.

Definition 2. (*Restricted Isometry Property*)

For all $\mathbf{x} \in \mathbb{R}^N$ *so that* $\| \mathbf{x} \|_0 \le k$, *if there exists* $0 \le \delta_k < 1$ *such that*

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \ \le \|\mathbf{A}\mathbf{x}\|_2^2 \le \quad (1 + \delta_k)\|\mathbf{x}\|_2^2 \tag{10}$$

is satisfied, then $\mathbf{A}$ *fulfills RIP of order k with radius* δ_k.

An equivalent description of the RIP is to say that all subsets of k columns taken from $\mathbf{A}$ are nearly orthogonal (the columns of $\mathbf{A}$ cannot be exactly orthogonal since we have more columns than rows) [16]. For example, if a matrix $\mathbf{A}$ satisfies the RIP of order $2k$, then we can interpret (10) saying that $\mathbf{A}$ approximately preserves the distance between any pair of k-sparse vectors. For random matrix $\mathbf{A}$, the following theorem is one of the results in relation to RIP for the noiseless CS problem, provided that the entries of the random matrix $\mathbf{A}$ are drawn from some distributions which are given later.

Theorem 1. (*Perfect Recovery Condition, Candes and Tao* [13])

If $\mathbf{A}$ *satisfies the RIP of order 2k with radius* δ_{2k}, *then for any k-sparse signal* $\mathbf{x}$ *sensed by* $\mathbf{y}=\mathbf{A}\mathbf{x}$, $\mathbf{x}$ *is with high probability perfectly recovered by the ideal program*

$$\widehat{\mathbf{x}} = arg \min_{\mathbf{x}} \quad \|\mathbf{x}\|_0$$
$$subject \ \ to \quad \mathbf{y} = \mathbf{A}\mathbf{x} \tag{11}$$

and it is unique, where $\| \mathbf{x} \|_0 = k \equiv \# \{i \in \{1, 2, \cdots, N\} | x_i \ne 0\}$.

This means, if $\mathbf{A}$ satisfies the RIP of order k with radius δ_k, then for any $k' < k$, $\mathbf{A}$ satisfies the RIP of order k' with constant $\delta_{k'} < \delta_k$ [?]. Note that, this theorem is stated for the noiseless CS problem and it is possible to extend it for the noisy CS system. The proof of these theorems is deferred to the literature mentioned, [13], in for the sake of space.

Under conventional sensing paradigm, the dimension of the original signal and the measurement should be at least equal. But in CS, the measurement vector can be far less than the original. While at the decoding side, reconstruction is done using nonlinear schemes. Eventually, the reconstruction is more cumbersome than the encoding which was only projections from a large space to a smaller space. On the other hand, finding a unique solution that satisfies the constraint that the signal itself is sparse or sparse in some domain is complex in nature. Fortunately, there are many algorithms to solve the CS problem, such as iterative methods such as *greedy iterative algorithms* [17] and *iterative thresholding algorithms* [18]. This chapter focuses merely on the *convex relaxations* methods [12, 13]. The regularizing terms in these methods can be reinterpreted as prior information under Bayesian inference. We consider a noisy measurement and apply convex relaxation algorithms for robust reconstruction.

2.3. Convex relaxation methods for CS

Various methods for estimating $\mathbf{x}$ may be used. We have the least square (LS) estimator in which no prior information is applied:

$$\widehat{\mathbf{x}} = \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{y}, \tag{12}$$

which performs very badly for the CS estimation problem we are considering. In order to incorporate the methods called convex relaxation, let us define an important concept first.

Definition 3. (*Unit Ball*)

A unit ball in l_p-space of dimension N can be defined as

$$\mathcal{B}_p \equiv \left\{ \mathbf{x} \in \mathbb{R}^N : \quad \|\mathbf{x}\|_p \leq 1 \right\}. \tag{13}$$

Unit balls corresponding to $p = 0$, $p = 1/2$, $p = 1$, $p = 2$, $p = \infty$, and $N = 2$, the balls are shown in **Figure 4**.

The exact solution for the noiseless CS problem is given by

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0, \text{ such that } \mathbf{y} = \mathbf{A}\mathbf{x}. \tag{14}$$

However, minimizing l_0-norm is a non-convex optimization problem which is NP-hard [19]. By relaxing the objective function to convexity, it is possible to get good approximation. That is, replacing the l_0-norm by the l_1-norm, one can find a problem which is tractable. Note that it is also possible to use other l_p-norms to relax the condition given by l_0. However, keeping our focus on l_1-norm, consider the minimization problem instead of (14).

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1, \text{ such that } \mathbf{y} = \mathbf{A}\mathbf{x} \tag{15}$$

The solution of the relaxed problem (15) gives the same as that of (14) and this equivalence was provided by Donoho and Huo in [20].
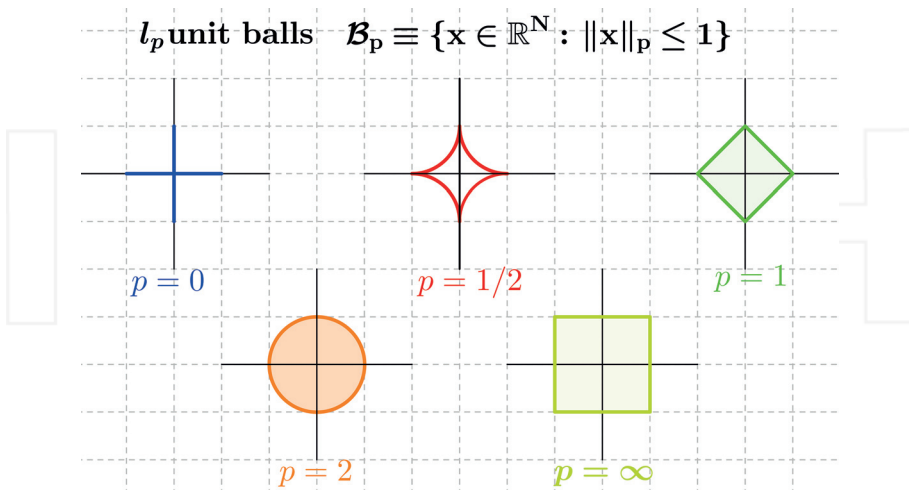
Figure 4. Different l_p-balls in different l_p-spaces for N=2, only balls with $p \geq 1$ are convex.

Theorem 2. (l_0-l_1 *Equivalence* [13])

If $\mathbf{A}$ *satisfies the RIP of order 2k with radius* $\delta_{2k} < \sqrt{2} - 1$, *then*

$$\widehat{\mathbf{x}} = arg\ \min_x \|\mathbf{x}\|_1$$
$$subject\ to\ \ \mathbf{y} = \mathbf{A}\mathbf{x} \tag{16}$$

is equivalent to (11) and will find the same unique $\widehat{\mathbf{x}}$.

Justified by this theorem, (15) is an optimization problem which can be solved in polynomial time and the fact that it gives the exact solution for the problem (14) under some circumstance has been one of the main reasons for the recent developments in CS. There is a simple geometric intuition on why such an approach gives good approximations. Among the l_p-norms that can be used in the construction of CS related optimization problems, only those which are convex give rise to a convex optimization problem which is more feasible than the non-convex counter parts, which means l_p-norms with only $p \geq 1$ satisfy such a condition. On the other hand, l_p-norms with $p > 1$ do not favor sparsity, for example, l_2-norm minimization tends to spread reconstruction across all coordinates even if the true solution is sparse. But l_1-norm is able to enforce sparsity. The intuition is that l_1-minimization solution is most likely to occur at corners or edges, not faces [21, 45]. That is why l_1-norm became famous for CS. Further, in CS literature, convex relaxation is presented as either l_2-penalized l_1-minimization called Basis Pursuit Denoising (BPDN) [22] or l_1-penalized l_2-minimization called least absolute shrinkage and selection operator (LASSO) [45], which are equivalent and effective in estimating a high-dimensional data.

Usually real world systems are contaminated with noise, $\mathbf{w}$, and in this chapter, the focus is on such problems. The noisy recovery problem becomes a simple extension of (15),

$$\min_\mathbf{x} \|\mathbf{x}\|_1, \text{such that } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \leq \epsilon \tag{17}$$

where ϵ is a bound on $\|\mathbf{w}\|_2$. The real problem for (17) is stability. Introducing small changes in the observations should result in small changes in the recovery. We can visualize this using the balls shown in **Figure 5**.

Both the l_0 and l_1-norms give exact solutions for the noise-free CS problem while giving a close solution for the noisy problem. However, the l_2-norm gives worst approximation in both cases compared to the other l_p-norms with $p < 2$ (see **Figure 5**). Moreover, (17) is equivalent to an unconstrained quadratic programming problem as

$$\min_x \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \gamma\|\mathbf{x}\|_1, \tag{18}$$

as it will be shown later as LASSO, where γ is a tuning parameter. The equivalency of (17) and (18) is shown in [23, 24]. In this chapter, the generalized form of the minimization problem in (18) with different l_p-norm regularization is considered, that is,
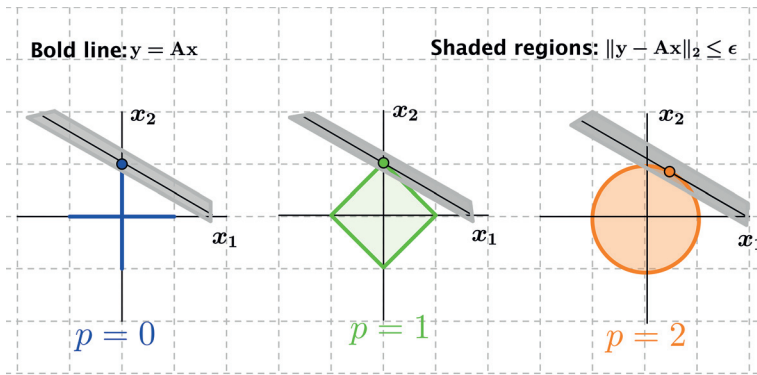
Figure 5. l_p-norm approximations: the constraints for the noise-free CS problem is given by the bold line while the shaded region is for the noisy one.

$$\min_x \frac{1}{2} \|\mathbf{y} - \mathbf{Ax}\|_2^2 + \gamma \|\mathbf{x}\|_p. \tag{19}$$

Further, this chapter provides the use of Bayesian framework in compressive sensing by incorporating two different priors modeling the sparsity and the possible structure among the sparse entries in a signal. Basically, it is the summary of the recent works [2, 25–27].

3. Bayesian inference used in CS problem

Under Bayesian inference, consider two random variables $\mathbf{x}$ and $\mathbf{y}$ with probability density function (pdf) $p(\mathbf{x})$ and $p(\mathbf{y})$, respectively. The product rule gives us $p(\mathbf{x},\mathbf{y}) = p(\mathbf{x}|\mathbf{y})p(\mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$ and Bayes' theorem provides

$$p(\mathbf{x}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{y})}. \tag{20}$$

Further, the maximum a posteriori (MAP), $\widehat{\mathbf{x}}_{MP}$, is defined as

$$\widehat{\mathbf{x}}_{MP} = \arg \max_{\mathbf{x}} \frac{p(\mathbf{y}|\mathbf{x})p(\mathbf{x})}{\int_{\tilde{\mathbf{x}}} p(\mathbf{y}|\tilde{\mathbf{x}})p(\tilde{\mathbf{x}})d\tilde{\mathbf{x}}} \tag{21}$$

$$= \arg \max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})p(\mathbf{x})$$

MAP is related to Fisher's methods of maximum likelihood estimation (MLE), $\widehat{\mathbf{x}}_{ML}$:

$$\widehat{\mathbf{x}}_{ML} = \arg \max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x}). \tag{22}$$

As we can see it from (21) and (22), the difference between MAP and MLE is the prior distribution. The former can be considered as a regularized form of the latter. Since we apply Bayesian inference we assume further different prior distributions on $\mathbf{x}$.

3.1. Sparse prior

The estimators of $\mathbf{x}$ resulting from (19) for the sparse problem we consider in this chapter, can be presented as a maximum a posteriori (MAP) estimator under the Bayesian framework as in [28]. We show this by defining a prior probability distribution for $\mathbf{x}$ on the form

$$p(\mathbf{x}) = \frac{e^{-uf(\mathbf{x})}}{\int_{\mathbf{x} \in \mathbb{R}^N} e^{-uf(\mathbf{x})} d\mathbf{x}} \tag{23}$$

where the regularizing function $f \colon \chi \to R$ is some scalar-valued, non negative function with $\chi \subseteq \mathbb{R}$ which can be expanded to a vector argument by

$$f(\mathbf{x}) = \sum_{i=1}^{N} f(x_i), \tag{24}$$

such that for sufficiently large u,

$$\int_{\mathbf{x} \in \mathbb{R}^N} \exp(-uf(\mathbf{x})) d\mathbf{x}$$

is finite. Furthermore, let the assumed variance of the noise be given by $\sigma^2 = \frac{\lambda}{u}$, where λ is the system parameter which can be taken as $\lambda = \sigma^2 u$.

Since the pdf of the noise $\mathbf{w}$ is gaussian, the likelihood function of $\mathbf{y}$ given $\mathbf{x}$ is given by

$$p_{\mathbf{y}|\mathbf{x}}(\mathbf{y}|\mathbf{x}) = \frac{1}{(2\pi\sigma)^{N/2}} e^{-\frac{1}{2\sigma^2}\|\mathbf{y} - A\mathbf{x}\|_2^2}. \tag{25}$$

Together with (20) and (23), this now gives

$$p_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}; \mathbf{A}) = \frac{e^{-u\left(\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x})\right)}}{(2\pi\sigma)^{N/2} \int_{\mathbf{x} \in \mathbb{R}^N} e^{-u\left(\frac{1}{2\lambda}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x})\right)} d\mathbf{x}}.$$

The MAP estimator, (21), becomes

$$\widehat{\mathbf{x}}_{MP} = \arg \min_{\mathbf{x} \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda f(\mathbf{x}). \tag{26}$$

Now, as we choose different regularizing function, we get different estimators as listed below [28]:

1. Linear estimators: when $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ (26) reduces to

$$\widehat{\mathbf{x}}_{\text{Linear}} = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T + \lambda \mathbf{I})^{-1}\mathbf{y}, \qquad (27)$$

which is the LMMSE estimator. But we ignore this estimator in our analysis since the results are not sparse. However, the following two estimators are more interesting for CS problems since they enforce sparsity into the vector $\mathbf{x}$.

2. LASSO estimator: when $f(\mathbf{x}) = \|\mathbf{x}\|_1$ we get the LASSO estimator and (26) becomes,

$$\hat{\mathbf{x}}\text{LASSO} = \arg\min_{x\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1. \qquad (28)$$

3. Zero-norm regularization estimator: when $f\mathbf{x} = \|\mathbf{x}\|_0$, we get the zero-norm regularization estimator and (26) becomes

$$\hat{\mathbf{x}}\text{Zero-Norm} = \arg\min_{x\in\mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_0. \qquad (29)$$

As mentioned earlier, (29) is the best solution for estimation of the sparse vector $\mathbf{x}$, but is NP-complete. The worst approximation for the sparse problem considered is the L2-regularization solution given by (27). However, the best approximation is given by Eq. (28) and its equivalent forms. We have used some of the algorithms in literature in our simulation which are considered as equivalent to this approximations such as Bayesian compressive sensing (BCS) [29] and L1-norm regularized least-squares (L1-LS) [11–13].

3.2. Clustering prior

The entries of the sparse vector $\mathbf{x}$ may have some special structure (clusteredness) among themselves. This can be modeled by modifying the previous prior distribution.[1] We use another penalizing parameter γ to represent clusteredness in the data. For that we define the clustering using the distance between the entries of the sparse vector $\mathbf{x}$ by

$$D \equiv \sum_{i=1}^{N} |x_i - x_{i-1}|,$$

and we use a regularizing parameter γ. Hence, we define the clustering prior to be

$$q(\mathbf{x}) = \frac{e^{-\gamma D(\mathbf{x})}}{\displaystyle\int_{\mathbf{x}\in\mathbb{R}^N} e^{-\gamma D(\mathbf{x})} d\mathbf{x}}. \qquad (30)$$

The new posterior involving this prior under the Bayesian framework is proportional to the product of the three pdfs:

---

[1] In [30] a hierarchical Bayesian generative model for sparse signals is found in which they have applied full Bayesian analysis by assuming prior distributions to each parameter appearing in the analysis. We follow a different approach.

$$p(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x})q(\mathbf{x}). \tag{31}$$

By similar argument s as used in 3.1, we arrive at the clustered LASSO estimator

$$\hat{\mathbf{x}}\text{Clu-Lasso} = \arg\min_{x \in \mathbb{R}^N} \frac{1}{2}\|\mathbf{y} - \mathbf{Ax}\|_2^2 + \lambda\|\mathbf{x}\|_1 + \gamma\sum_{i=1}^{N}|x_i - x_{i-1}|. \tag{32}$$

Here λ, γ are our tuning parameters for the sparsity in $\mathbf{x}$ and the way the entries are clustered, respectively.

4. Bayesian inference in CS applications

Compressed sensing paradigm has been applied to many signal processing areas [31–41]. However, at this time, building the hardware that can translate the CS theory into practical use is very limited. Nonetheless, the demand for cheaper, faster, and efficient devices will motivate the use of CS paradigm in real-time systems in the near future.

So far, in image processing, one can mention the single-pixel imaging via compressive sampling [31], in magnetic resonance imaging (MRI) for reducing scan time and improved image quality [32], in seismic images [33], and in radar systems for simplifying hardware design and to obtain high resolution [34, 35]. In communication and networks, CS theory has been studied for sparse channel estimation [36], for under water acoustic channels which are inherently sparse [37], spectrum sensing in cognitive radio networks [38], for large wireless sensor networks (WSNs) [39], as a channel coding scheme [40], localization [41] and so on. A good CS application literature review is provided in [21], which basically is the summary of the bulk of literatures given at http://dsp.rice.edu/cs.

In this chapter, there are examples of CS theory applications using Bayesian inference in imaging, like magnetic resonance imaging (MRI) and, in communication, i.e., multiple-input multiple-output (MIMO) systems, and in remote sensing. First, let us see the impact of the estimators derived above, LASSO and clustered LASSO, in MRI.

4.1. Magnetic resonance imaging (MRI)

MRI images are usually very weak due to the presence of noise and due to the weak nature of the signal itself. Compressed sensing (CS) paradigm can be applied in order to boost such signal recoveries. We applied CS paradigm via Bayesian framework, that is, incorporating the different prior information such as sparsity and the special structure that can be found in such sparse signal recovery method is applied on different MRI images.

4.1.1. Angiogram image

Angiogram images are already sparse in the pixel representation. An angiogram image taken from the University Hospital Rechts der Isar, Munich, Germany [42] is used for our analysis.
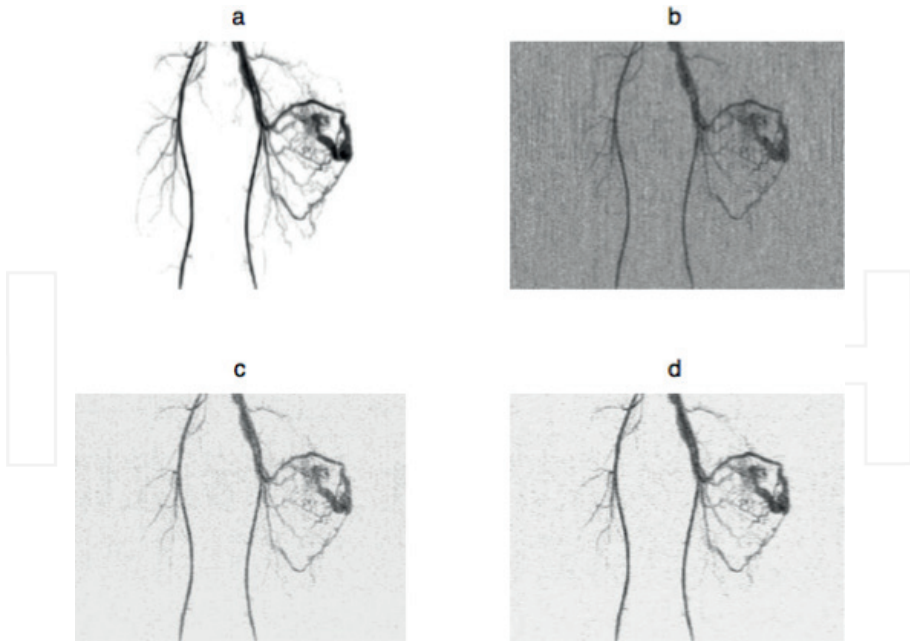
Figure 6. Comparison of reconstruction schemes together with performance comparison using mean square error (MSE) in dB: (a) original image $\mathbf{x}$; (b) LMMSE (-35.1988 dB); (c) LASSO (-53.6195 dB); and (d) clustered Lasso (-63.6889 dB).

The image we took is sparse and clustered even in the pixel domain. The original signal after vectorization is $\mathbf{x}$ of length $N = 960$. By taking 746 measurements, and maximum number of non-zero elements $k = 373$, we applied different reconstruction schemes and the results are shown in **Figure 6**.

4.1.2. Phantom image

Another MRI image considered is the Shepp-Logan phantom which is not sparse in spatial domain. However, we sparsified it in K-space by zeroing out small coefficients. We then measured the sparsified image and added noise. The original signal after vectorization is $\mathbf{x}$ of length $N = 200$. By taking 94 measurements, that is, $\mathbf{y}$ is of length $M = 94$, and maximum number of non-zero elements $k = 47$, we applied different reconstruction algorithms used above. The result shows that clustered LASSO does well compared to the others as can be seen in **Figure 7**.

4.1.3. fMRI image

Another example to apply the clustered LASSO based image reconstruction using Bayesian framework to medical images is a functional magnetic resonance imaging (fMRI), a non-invasive technique of brain mapping, which is crucial in the study of brain activity. Taking many slices in fMRI data, we saw how these data sets are sparse in the Fourier domain. This is
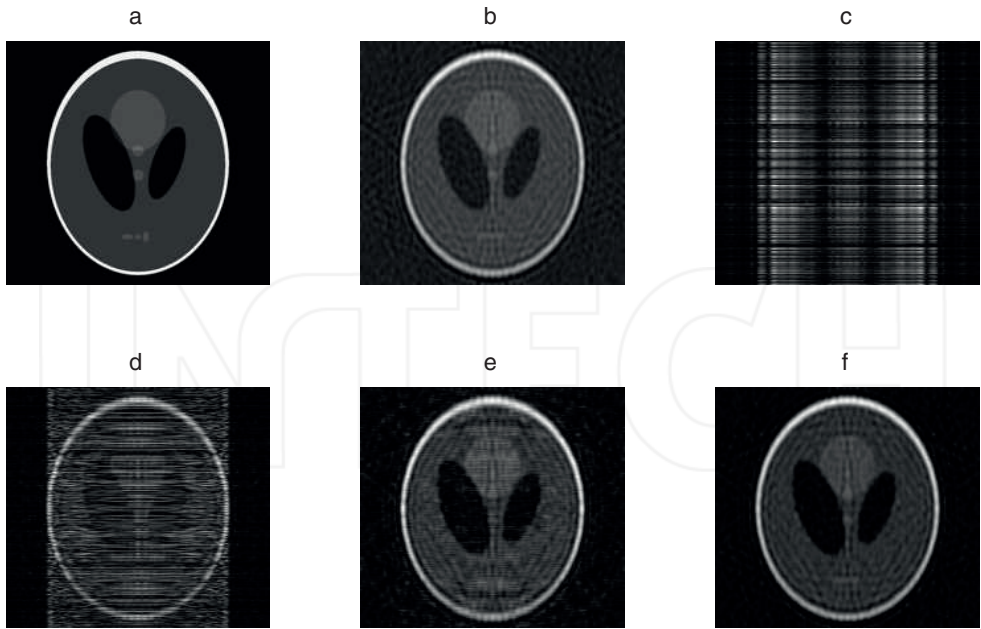
Figure 7. Comparison of reconstruction schemes together with performance comparison using mean square error (MSE) in dB: (a) original image x; (b) sparcified image; (c) least square (LS) (-21.3304 dB); (d) LMMSE (-27.387 dB); (e) LASSO (-37.9978 dB); and (f) clustered LASSO (-40.0068 dB).
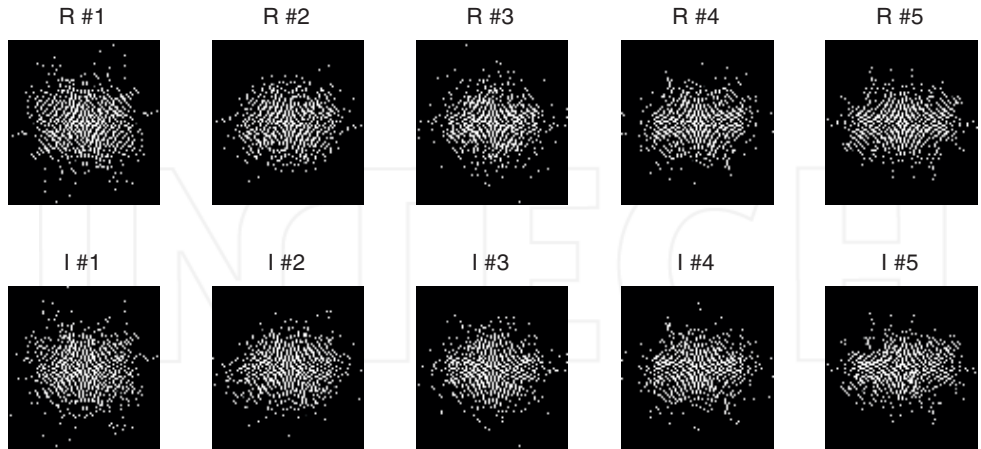
Figure 8. The five column images represent the real and imaginary parts of the Fourier transform representation of the data set we have chosen to present further, which in general shows that the fMRI image have sparse and clustered representation.

shown in **Figure 8**. We observed the whole data in this domain for the whole brain image. They all share the characteristics we have based our analysis, i.e., sparsity and clusteredness. Then we took some slices which are consecutive in the slice order and took different N, k, and $M=2k$, on these slices. We can see the numbers at the top of **Figure 9**, in which the two numbers represent k and N, respectively.

In fMRI, results are compared using image intensity which gives a good ground for a health practitioner to observe and decide in accordance to the available information. The more one have prior knowledge on how the brain regions work in human beings or pets the better priors that one incorporate to analyze the data. So this is an interesting tool for researchers in the future.

4.2. MIMO systems

Multiple-input multiple-output (MIMO) systems are integrated in modern wireless communications due to their advantage in improving performance with respect to many performance metrics. One such advantage is the ability to transmit multiple streams using spatial multiplexing, but channel state information (CSI) at the transmitter is needed to get optimal system performance.

Consider a frequency division duplex (FDD) MIMO system consisting of N_t transmit and N_r receive antennas. Assume that the channel is a flat-fading, temporally correlated channel denoted by a matrix $\mathbf{H}[n] \in \mathbb{C}^{N_r \times N_t}$ where n indicates a channel feedback time index with block fading assumed during the feedback interval. The singular value decomposition (SVD) of $\mathbf{H}[n]$ gives

$$\mathbf{H}[n] = \mathbf{U}[n]\, \Sigma\,[n]\mathbf{V}^H[n],$$

where $\mathbf{U} \in \mathbb{C}^{N_r \times r}$ and $\mathbf{V} \in \mathbb{C}^{N_t \times r}$ are unitary matrices and $\boldsymbol{\Sigma} \in \mathbb{C}^{r \times r}$ is a diagonal matrix consisting of $r = \min(N_t, N_r)$ singular values. In the presence of perfect channel state information (CSI), a MIMO system model can be given by the equation

$$\tilde{\mathbf{y}} = \mathbf{U}^H[n]\mathbf{H}[n]\mathbf{V}[n]\tilde{\mathbf{x}} + \mathbf{U}^H[n]\mathbf{n} \tag{33}$$

where $\tilde{\mathbf{x}} \in \mathbb{C}^{r \times 1}$ is transmitted vector, $\mathbf{V}[n]$ is used as precoder at the transmitter, $\mathbf{U}^H[n]$ is used as decoder at the receiver, $\mathbf{n} \in \mathbb{C}^{N_r \times 1}$ denotes a noise vector whose entries are i.i.d. and distributed according to $\mathcal{CN}(0, 1)$ and $\tilde{\mathbf{y}} \in \mathbb{C}^{N_r \times 1}$ is the received vector.

Channel adaptive transmission requires knowledge of channel state information at the transmitter. In temporally correlated MIMO channels, the correlation can be utilized to reduce feedback overhead and improve performance. CS methods and rotative quantization are used to compress and feedback the CSI for MIMO systems [43]. This was done as an extension work of [44]. It is shown that the CS-based method reduces feedback overhead while delivering the same performance as the direct quantization scheme, using simulation.

Three methods are compared in the simulations, perfect CSI, without CS and with CS using matched filter (MF) and minimum mean square error estimator (MMSE) receivers for different

total feedback bits $B = 10$ and $B = 5$. In **Figure 10**, sum rates are compared against signal-to-noise-ratio (SNR). Using CS, half of the number of bits can be saved. In **Figure 11**, where the bit-error-rate is plotted against SNR, the CS method has a better bit error rate performance using same number of bits for the CS and without CS cases. These two figures demonstrate the clear advantage of using CS in feedback of singular vectors in rotative based method. The detail is deferred to [43].

4.3. Remote sensing

Remote sensing satellites provide a repetitive and consistent view of the Earth and they offer a wide range of spatial, spectral, radiometric, and temporal resolutions. Image fusion is applied to extract all the important features from various input images. These images are integrated to form a fused image which is more informative and suitable for human visual perception or computer processing. Sparse representation has been applied to fuse image to improve the quality of fused image [45].
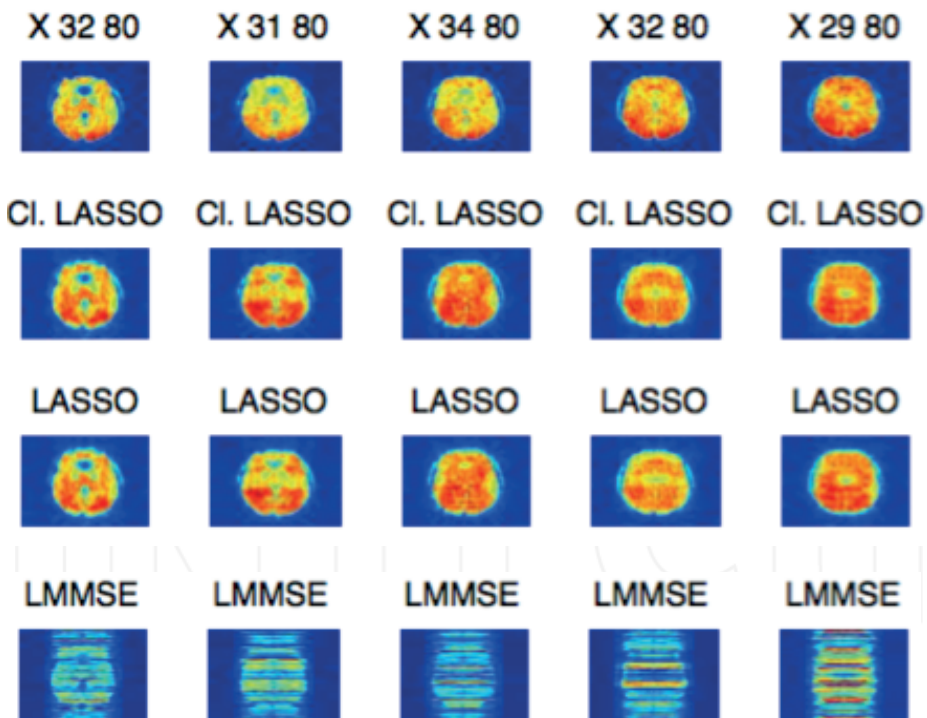
Figure 9. Application of sparse and cluster prior, LASSO and clustered LASSO (CL. LASSO), on a fMRI data analysis for $N = 80$, $k > 50$ for $\sigma^2 = 0.1$ and $\lambda = 0.1$, where LMMSE is with L2-regularised one.
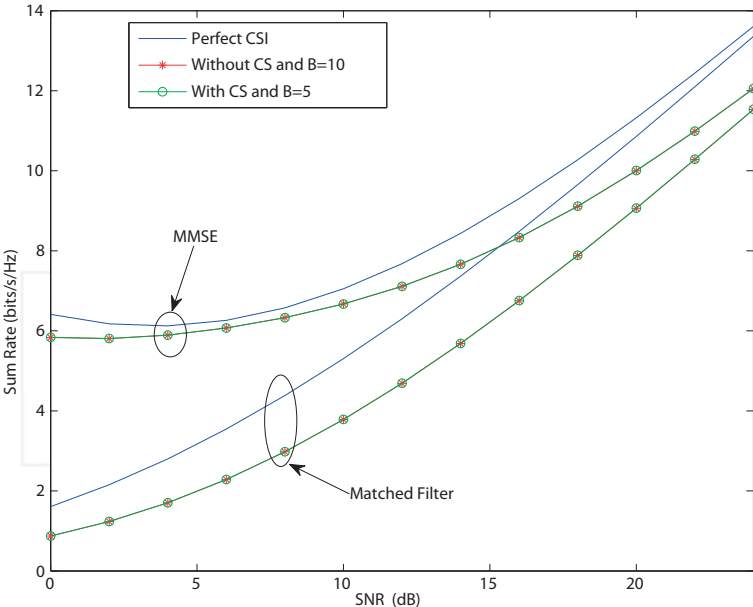
Figure 10. Sum rate vs. SNR for a 2×2 MIMO system with and without CS with two streams. We can observe that the performance of the CS method is almost equal to that of the method without using CS while saving half the number of bits.
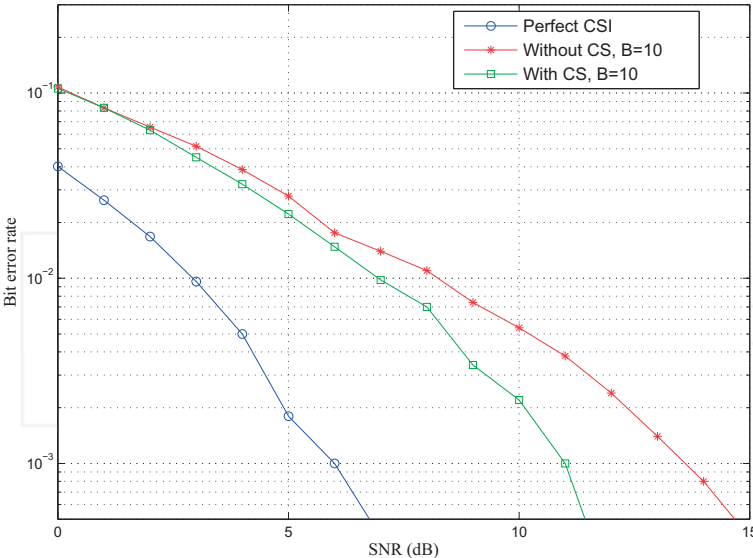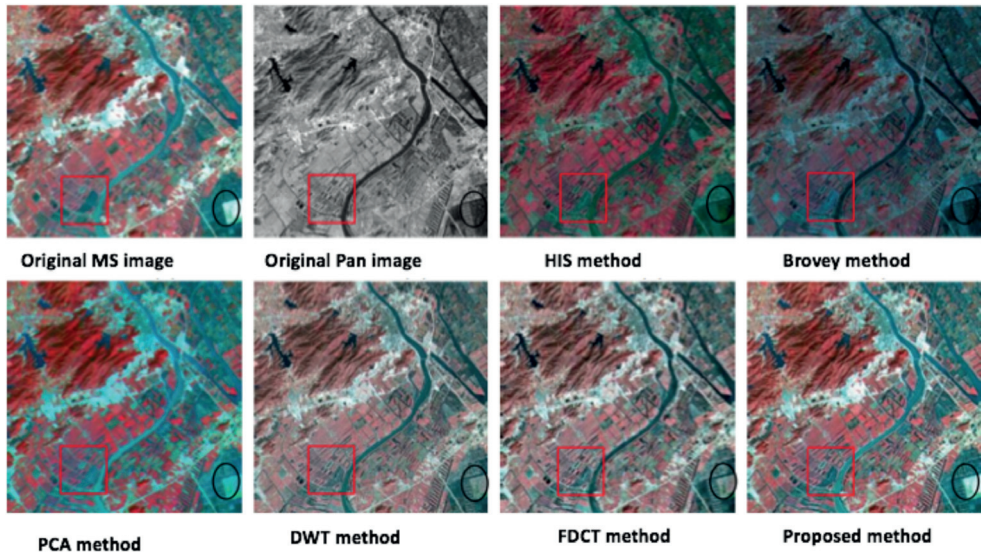

Figure 11. Bit error rate vs. SNR using matched filter receiver for a 2×2 MIMO system with one stream.

Figure 12. Comparison of image fusion methods for remote sensing applications using Brovey, DWT, PCA, FDCT, and the sparse representation methods [46].

To improve the quality of the fused image, a remote sensing image fusion method based on sparse representation is proposed in [46]. In these methods, the source images were represented with sparse coefficients first. Then, the larger values of sparse coefficients of panchromatic (Pan) image are set to 0. Thereafter, the coefficients of panchromatic (Pan) and multispectral (MS) image are combined with the linear weighted averaging fusion rule. Finally, the fused image is reconstructed from the combined sparse coefficients and the dictionary. The proposed method is compared with intensity-hue-saturation (IHS), Brovey transform (Brovey), discrete wavelet transform (DWT), principal component analysis (PCA) and fast discrete curvelet transform (FDCT) methods on several pairs of multifocus images. The proposed method using sparse representation outperforms, see **Figure 12**, better than the usual methods listed here. We believe that our method of clustered compressed sensing can also further improve this result.

5. Conclusions

In this chapter, a Bayesian way of analyzing data on CS paradigm is presented. The method assumes prior information like the sparsity and clusteredness of signals in the analysis of the data. Among the different reconstruction methods, the convex relaxation methods are redefined using Bayesian inference. Further, three CS applications are presented: MRI imaging, MIMO systems, and remote sensing. For MRI imaging, the two different priors are incorporated, while for MIMO systems and remote sensing, only the sparse prior is applied in

the analysis. We suggest that including the special structure among the sparse elements of the data can be included in the analysis to further improve the results.

Author details

Solomon A. Tesfamicael[1]* and Faraz Barzideh[2]

*Address all correspondence to: solomon.a.tesfamicael@ntnu.no

1 Department of Education, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

2 Department of Electrical Engineering and Computer Science, University of Stavanger (UiS), Stavanger, Norway

References

[1] Jaynes ET. Probability Theory: The Logic of Science. Cambridge University Press; 2003

[2] Tesfamicael SA, Barzideh F. Clustered compressed sensing in fMRI data analysis using a Bayesian framework. International Journal of Information and Electronics Engineering. 2014;**4**(2):74-80

[3] Mackay DJC. Information Theory, Inference, and Learning Algorithms. Cambridge University Press; 2003. ISBN: 978-0-521-64298-9

[4] O'Hagen A, Forster J. Kendall's Advanced Theory of statistics, volume 2B. Bayesian Inference. Arnold, a member of the Hodder Headline Group; 2004. ISBN: 0 340 807520

[5] Berger JO. Bayesian and Conditional Frequentist Hypothesis Testing and Model Selection. VIII C:L:A:P:E:M; La Havana, Cuba; November 2001

[6] Efron B. Modern Science and the Bayesian-Frequentist Controversy; 2005-19B/233. January 2005

[7] Botje MRA. Fisher on Bayes and Bayes' Theorem, Bayesian Analysis; 2008

[8] Moreno E, Javier Giron F. On the Frequentist and Bayesian approaches to hypothesis testing. January-June 2006;3-28

[9] Friston KJ, Penny W, Phillips C, Kiebel S, Hinton G, Ashburner J, Classical and Bayesian inference in neuroimaging: Theory. NeuroImage. June 2002;**16**(2):465-483

[10] Donoho D. Compressed sensing. IEEE Transactions on Information Theory. 2006;**52**(4): 12891306

[11] Candes EJ, Tao T. Decoding by linear programming. IEEE Transactions on Information Theory. December 2005;**51**(12)

[12] Cand'es E, Romberg J, Tao T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. IEEE Transactions on Information Theory. February 2006;**52**(2):489509

[13] Cande's EJ, Tao T. Near-optimal signal recovery from random projections: Universal encoding strategies? IEEE Transactions on Information Theory. December 2006;**52**:54065425

[14] Eldar YC, Kutyniok G. Compressed Sensing: Theory and Applications. Cambridge University Press; 2012

[15] Eldar YC, Kutyniok G. Compressed Sensing: Algorithms and Applications. KTHKTH, Communication Theory. ACCESS Linnaeus Centre; 2012

[16] Candes EJ. The restricted isometry property and its implications for compressed sensing. Comptes Rendus Mathematique. 2008

[17] Guan X, Gao Y, Chang J, Zhang Z. Advances in Theory of Compressive Sensing and Applications in Communication. 2011 First International Conference on Instrumentation, Measurement, Computer, Communication and Control; 2011

[18] Blumensath T, Davies ME. Iterative hard thresholding for compressed sensing. Applied and Computational Harmonic Analysis. 2009

[19] Natarajan BK. Sparse approximate solutions to linear systems. SIAM Journal of Computing. 1995

[20] Natarajan BK. Uncertainty principles and ideal atomic decomposition. IEEE Transactions on Information Theory. January 2001;**47**:2845-2862

[21] Qaisar S, Bilal RM, Iqbal W, Naureen M, Lee S. Compressive sensing: From theory to applications, a survey. Journal of Communications and Networks. October 2013;**15**(5):443-456

[22] Figueiredo MAT, Nowak RD, Wright SJ. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. Journal of Selected Topics in Signal Processing. 2007;**1**(4):586-597

[23] Schniter P, Potter LC, Ziniel J. Subspace pursuit for compressive sensing signal reconstruction. Information Theory and Applications Workshop. February 2008. pp. 326-333

[24] Teixeira FCA, Bergen SWA, Antoniou A. Robust signal recovery approach for compressive sensing using unconstrained optimization. Proceedings of 2010 IEEE International Symposium on Circuits and Systems (ISCAS). May 2010. pp. 3521-3524

[25] Tesfamicael SA, Barzideh F. Clustered compressed sensing via Bayesian framework. IEEE UKSim-AMSS 17th International Conference on Computer Modelling and Simulation. UKSim2015-19.S.Image, Speech and Signal Processing; Cambridge, United Kingdom. 2015. pp. 25-27

[26] Tesfamicael SA, Barzideh F. Clustered compressive sensing: Application on medical imaging. International Journal of Information and Electronics Engineering. 2015;**5**(1):48-50

[27] Tesfamicael SA. Compressive sensing in signal processing: Performance analysis and applications [doctoral thesis]. NTNU; 2016. p. 182

[28] Rangan S, Fletcher AK, Goyal VK. Asymptotic Analysis of MAP Estimation via the Replica Method and Applications to Compressed Sensing. 2009. arXiv:0906.3234v1

[29] Ji S, Xue Y, Carin L. Bayesian compressive sensing. IEEE Transactions on Signal Processing. June 2008;**56**(6):2346-2356

[30] Yu L, Sun H, Pierre Barbot J, Zheng G. Bayesian compressive sensing for clustered sparse signals. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2011

[31] Duarte MF, Davenport MA, Takhar D, Laska JN, Sun T, Kelly KF, Baraniuk RG. Single-pixel imaging via compressive sampling. IEEE Signal Processing Magazine. March 2008;**25**(2):83-91

[32] Lustig M, Donoho D, Pauly JM. Sparse MRI: The application of compressed sensing for rapid MR imaging. Magnetic Resonance in Medicine. 2007;**58**(6):1182-1195

[33] Lustig M, Donoho D, Pauly JM. Simply denoise: Wavefield reconstruction via jittered undersampling. Geophysics. 2007;128-133

[34] Baraniuk R, Steeghs P. 2007 IEEE Compressive Radar Imaging Radar Conference. 19-28 April, 2007

[35] Herman MA, Strohmer T. High-resolution radar via compressed sensing. IEEE Transactions on Signal Processing. June 2009;**57**(6):2275-2284

[36] Bajwa WU, Haupt J, Sayeed AM, Nowak R. Compressed channel sensing: A new approach to estimating sparse multipath channel. Proceedings of the IEEE. June 2010;**98**(6):1058-1107

[37] Berger CR, Zhou S, Preisig JC, Willett P. Sparse channel estimation for multicarrier underwater acoustic communication: From subspace methods to compressed sensing. IEEE Transactions on Signal Processing. March 2010;**58**(3):1708-1721

[38] Zeng F, Zhi T, Chen L. Distributed compressive wideband spectrum sensing in cooperative multi-hop cognitive networks. IEEE International Conference on Communications (ICC). 1-5 May, 2010

[39] Qing L, Zhi T. Decentralized sparse signal recovery for compressive sleeping wireless sensor networks. IEEE Transactions on Signal Processing. July 2010;**58**(7):3816-3827

[40] Charbiwala Z, Chakraborty S, Zahedi S, Kim Y, Srivastava MB, He T, Bisdikian C. Compressive oversampling for robust data transmission in sensor networks. IEEE INFOCOM Proceedings. 2010. pp. 1-9

[41] Chen F, Au WSA, Valaee S, Zhenhui T. Compressive sensing based positioning using RSS of WLAN access points. IEEE INFOCOM Proceedings. 1-9 March 2010

[42] Image. Depiction of Vessel Diseases with a Wide Range of Contrast and Non-Contrast Enhanced Techniques. Munich, Germany: University Hospital Rechts der Isar; 2014

[43] Tesfamicael SA, Lundheim L. Compressed sensing based rotative quantization in temporally correlated MIMO channels. Recent Developments in Signal Processing; 2013

[44] Godana SBE, Ekman T. Rotative quantization using adaptive range for temporally correlated MIMO channels. 2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC). 2013. pp. 1233-1238

[45] Tibshirani R. Compressive sensing: From theory to applications, a survey. Journal of the Royal Statistical Society. Series B (Methodological). 1996;**58**(1):267-288

[46] Yu X, Gao G, Xu J, Wang G. Remote sensing image fusion based on sparse representation. 2014 IEEE Geoscience and Remote Sensing Symposium