Chapter 3
Compressed Sensing with Side Information on Feasible Region

In the literature the problem of compressed sensing in the presence of side information is studied. But, in most cases the side information are about the source itself, i.e. structure, probability distribution, etc. In this chapter, the problem of compressed sensing in the presence of side information about the feasible region is reviewed. We follow an approach similar to [1] to formulate the problem mathematically for a wider class. Next it is shown that uniqueness and stability results of CS still holds in this formulation. Finally, an efficient recovery algorithm is derived which incorporates the side information.

3.1 Formulation

Consider a general compressed sensing problem (1.3). Assume null(A) satisfies spherical section property with parameter Δ, consequently Theorems 2.3.1 and 2.3.2 hold for this problem. From linear algebra if $\mathbf{s}_1$ is a special solution to the system $A\mathbf{s} = \mathbf{y}$, then the feasible region for the optimization problem (1.3) would be:

$$F = \{\mathbf{s}_1 + \mathbf{s}_2 | A\mathbf{s}_2 = 0\} = \{\mathbf{s}_1\} + \text{null(A)}. \tag{3.1}$$

In the current work we adopt FISTA as the reconstruction algorithm. To solve for the unique solution we start from an initial point and then search in the feasible region in (2.34). The size of this region depends on null(A) $\equiv \mathbb{R}^{\text{m}}$ $(rank(A))$ and intuitively we expect the bigger this space is, the harder is to solve the optimization problem. In other words when the feasible region is small then (2.34) converges faster to the solution of (1.4). Thus any side information about the feasible region is helpful.

Now consider cases that we have side information about the feasible region. For instance in the case of derivative compressed sensing (DCS) [1], where the source signal is a gradient field, the side information will result in $B\mathbf{s} = 0$ condition on the source signal ($B \in \mathbb{R}^{\frac{n}{2} \times n}$ and is resulted from inherent property of a gradient field.

M. Rostami, *Compressed Sensing with Side Information on the Feasible Region*,
SpringerBriefs in Electrical and Computer Engineering,
DOI: 10.1007/978-3-319-00366-5_3, © The Author(s) 2013

We will discuss this special case in more details in Chap. 4). A more general case may happen when we have side information as:

$$B\mathbf{s} = \mathbf{b}, \tag{3.2}$$

where $B \in \mathbb{R}^{m' \times n}$ is a full rank matrix. Many constraints on a source can be formulated as (3.2). In such cases we have two types of information about the source. We call the first type, primary information, which is resulted through measurements ($A\mathbf{s} = \mathbf{y}$). The secondary information comes in hand through an inherent property of the source. Broadly speaking, we can assume we have a general inverse problem ($B\mathbf{s} = \mathbf{b}$) and we also have sparsity prior on the source, then we apply CS as a regularization method on this problem. Some problems in image/signal processing area such as image super-resolution, image impainting, and medical imaging can be modeled in this framework. We expect that if we incorporate this side information it somehow improves CS reconstruction. For instance we may be able to decrease the number of measurements for recovering the source with similar accuracy or some kind of robustness towards noise when the measurements are noisy.

Let $A' = \begin{bmatrix} A \\ B \end{bmatrix}$ ($A' \in \mathbb{R}^{(m+m') \times n}$ and is full rank matrix) and $\mathbf{y}' = \begin{bmatrix} \mathbf{y} \\ \mathbf{b} \end{bmatrix}$ ($\mathbf{y}' \in \mathbb{R}^{(m+m')}$). We have the following equivalent problem:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \|\mathbf{s}\|_0 \quad \text{s.t.} \quad \mathbf{y}' = A'\mathbf{s}. \tag{3.3}$$

Assume $m + m' \leq n$, in such cases the new problem is exactly in theform of a CS problem. We assume $m + m' \leq n$ so as to ensure an underdetermined system to deal with problem in CS framework. Now the question is: Does this problem have a unique solution? Can we still replace the l_0-norm with l_1-norm? To answer these questions, the new sensing matrix A' must be studied. In the next section we will show the answers to both questions are positive.

3.2 Uniqueness and Stability

Consider the optimization problem (3.3). Our assumption is that A has Δ-spherical section property, so (1.3) has a unique solution $\hat{\mathbf{s}}$ and also we have l_1-l_0 equivalence. We show that adding the secondary condition will not violate the uniqueness, and furthermore solutions of (1.3) and (3.3) are equal.

Lemma 1 *The problem* (3.3) *has a unique solution* $\hat{\mathbf{s}}$ *equivalent to the solution of* (1.3). *Furthermore* l_1-l_0 *equivalence holds for this problem, i.e.:*

$$\hat{\boldsymbol{s}} = \arg\min_{\boldsymbol{s}} \|\boldsymbol{s}\|_1 \quad s.t. \quad \boldsymbol{y}' = A'\boldsymbol{s}. \tag{3.4}$$

Proof 3 *The proof is simple. First we show* A' *has spherical section property with* $\Delta' = (1 + \frac{m'}{m})\Delta$. *Note dim*(null(A′)) = n − (m + m′) *and:*

$$A's = 0 \rightarrow \begin{cases} As = 0 \\ Bs = 0 \end{cases} \rightarrow \text{null}(A') = \text{null}(A) \cap \text{null}(B), \tag{3.5}$$

thus null(A′) ⊂ null(A). *Consequently* $\forall \boldsymbol{s} \in$ null(A′) ⊂ null(A)*:*

$$\frac{\|\boldsymbol{s}\|_1}{\|\boldsymbol{s}\|_2} \geq \sqrt{\frac{m}{\Delta}} = \sqrt{\frac{m+m'}{\Delta(1+\frac{m'}{m})}}, \tag{3.6}$$

according to the definition of SSP, null(A′) *has spherical section property with* $\Delta = (1 + \frac{m'}{m})\Delta$.

According to the assumption $\hat{\boldsymbol{s}}$ *is the solution of* (1.3) *and also satisfies* $B\boldsymbol{s} = \boldsymbol{b}$, $\hat{\boldsymbol{s}}$ *lays in the feasible region of* (3.3). *According to Theorem 2.3.1,* $\hat{\boldsymbol{s}}$ *is a unique solution of* (3.3), *if* $\|\hat{\boldsymbol{s}}\|_0 \leq \frac{m+m'}{2\Delta'}$. *Note* $\|\hat{\boldsymbol{s}}\|_0 \leq \frac{m}{2\Delta}$ *since it is the unique solution of* (1.3), *then from Theorem 2.3.1:*

$$\|\hat{\boldsymbol{s}}\|_0 \leq \frac{m}{2\Delta} = \frac{m(m+m')}{2\Delta(m+m')} = \frac{m+m'}{2\Delta(1+\frac{m'}{m})} = \frac{m+m'}{2\Delta'}, \tag{3.7}$$

which concludes the proof. Similarly we can conclude if the original primary CS problem has l_1-l_0 *equivalence in Theorem 2.3.2, then* (3.3) *inherits this property. Also note that this unique solution satisfies* $A\boldsymbol{s} = \boldsymbol{y}$ *and thus is equal to solution of* (1.3).

This lemma states that we can add any side information in the form of (3.2) to our problem and this will not make the situation worse. This result is very intuitive and is expected but the Lemma also gives a mathematical justification. For source reconstruction we can use a general proposed CS reconstruction algorithm and find the unique solution of (3.3), however this may not be efficient enough. In the next section, a more efficient algorithm is proposed to solve (3.3).

3.3 Numerical Solution Algorithm

As explained, the problem that we formulated in Sect. 3.1 can be formulated by (3.3). We also expect some improvement if we use side information. In this section an efficient algorithm is derived for solving this problem. When Lemma 1 holds, (3.3) can be equivalently formulated as follows:

$$\hat{\mathbf{s}} = \arg\min_{\mathbf{s}} \mu\|\mathbf{s}\|_1 + \|A\mathbf{s} - \mathbf{y}\|_2^2 \quad \text{s.t.} \quad B\mathbf{s} = \mathbf{b}, \tag{3.8}$$

now we have our original CS reconstruction problem constrained to the side information. To solve optimization problems in this form one can use operator splitting [2–4]. We will have a quick review on this method and then use it to solve (3.8).

3.3.1 Bregman Iteration and Operator Splitting

Consider the following optimization problem:

$$\min_{\mathbf{s}} J(\mathbf{s}) \quad \text{s.t.} \quad H(\mathbf{s}) = 0, \tag{3.9}$$

where H is a convex differentiable furcation while J is also convex but possibly non-differentiable functions. An efficient method to solve this type of problems is to use the Bregman iterations [2].

To proceed we need the definition of sub-gradient and Bregman distance.

Definition 6 *Let $J(\cdot) : \mathbb{R}^n \to \mathbb{R}^+$ be a convex and possibly non-differentiable function. The vector $\boldsymbol{p} \in \mathbb{R}^n$ is called a sub-gradient of J at point $\boldsymbol{w}_0$:*

$$\forall \boldsymbol{w} \in \mathbb{R}^n : J(\boldsymbol{w}) - J(\boldsymbol{w}_0) \leq \langle \boldsymbol{p}, \boldsymbol{w} - \boldsymbol{w}_0 \rangle. \tag{3.10}$$

Also, the set of all $\boldsymbol{p}$'s is called sub-differentiable of J at point $\boldsymbol{w}$ and is denoted by $\partial J(\boldsymbol{w}_0)$.

For a differentiable function, $\partial J(\mathbf{w}_0)$ reduces to a singleton which only contains the gradient vector, $\nabla J(\mathbf{w}_0)$. This concept extends the definition of gradient to convex but possibly non-differentiable functions. For instance sub-differentiable of $J(w) = |w|$ at the point $w = 0$ is the set $[-1, 1]$. Next we require definition of Bregman distance

Definition 7 *The Bregman distance of a convex function $J(\cdot) : \mathbb{R}^n \to \mathbb{R}^+$ between two points $\boldsymbol{s}$ and $\boldsymbol{w}$ is defined as:*

$$D_J^{\boldsymbol{p}}(\boldsymbol{s}, \boldsymbol{w}) = J(\boldsymbol{s}) - J(\boldsymbol{w}) - \langle \boldsymbol{p}, \boldsymbol{s} - \boldsymbol{w} \rangle, \tag{3.11}$$

where $\boldsymbol{p}$ is a sub-gradient of J at $\boldsymbol{w}$.

Note (3.11) is not symmetric, thus Bregman distance is not a metric but somehow measures closeness of the two points.

Now back in our problem (3.9), this problem can be solved iteratively as follows:

$$\begin{cases} \mathbf{s}^{i+1} = \arg\min_{\mathbf{s}} D_J^{\mathbf{p}^i}(\mathbf{s}, \mathbf{s}^i) + \delta H(\mathbf{s}) \\ \mathbf{p}^{i+1} = \mathbf{p}^i - \delta \partial H(\mathbf{s}^{i+1}), \end{cases} \tag{3.12}$$

where $\delta \geq 0$ is a constant. It is shown in [4], that if the original problem (3.9) has a solution $\hat{\mathbf{s}}$, then through the iterations in (3.12), as $i \to \infty$ then $\mathbf{s}^i \to \hat{\mathbf{s}}$.

Now we apply this algorithm on (3.8), for which we can assume, $H(\mathbf{s}) = \frac{1}{2}\| B\mathbf{s} - \mathbf{b}\|_2^2$ and $J(s) = \mu\|\mathbf{s}\|_1 + \|A\mathbf{s} - \mathbf{y}\|_2^2$. This will reduce (3.12):

$$\begin{cases} (\mathbf{s}^{i+1}, \mathbf{b}^{i+1}) = \arg\min_{\mathbf{s},\mathbf{b}} \mu\|\mathbf{s}\|_1 + \frac{1}{2}\|A\mathbf{s} - \mathbf{y}\|_2^2 + \frac{\delta}{2}\|B\mathbf{s} - \mathbf{b} + \mathbf{p}^i\|_2^2 \\ \mathbf{p}^{i+1} = \mathbf{p}^i + \delta B\mathbf{s}^{i+1} - \mathbf{b}^{i+1}. \end{cases} \tag{3.13}$$

Note that the update step of the first equation in (3.13) has the format of a standard basis pursuit de-noising (BPDN) problem [5], which can be solved by a variety of optimization methods [6]. In the present paper, we used the FISTA algorithm of [7] due to the simplicity of its implementation as well as for its remarkable convergence properties. It should be noted that the algorithm does not require explicitly defining the matrices A and B. Only the *operations* of multiplication by these matrices and their transposes need to be known, which can be implemented in an implicit and computationally efficient manner. The main advantage of solving the problem using operator splitting is the much faster convergence of the thresholding algorithm.

Now equipped with some theoretical evidence and an efficient reconstruction algorithm we continue with some experimental study on advantageous prospect of our approach.

3.4 Experimental Study

To verify our analysis and algorithm, this section is devoted to experimental study on synthetics data, where as in the next Chapters we will focus on the practical applications of the developed method.

3.4.1 Source Model

For source simulation we used mixture of Gaussian model as the sparse source model:

$$\mathbf{s} \sim pN(0, \sigma_1) + (1 - p)N(0, \sigma_2), \tag{3.14}$$

where $N(0, \sigma_i)$ denotes a Gaussian distribution with zero mean and variance σ_i^2, $\sigma_1 \ll \sigma_2$, and p is the parameter for a Bernoulli distribution. This model has been used to represent sparse signals in the literature [8, 9]. Although it is not a proper model for some applications, it is useful for our experimental study. Here, it is assumed that the source signal has two states. The first state corresponds to source elements with large values (non-zero elements) and the second state corresponds to elements with negligible value (approximately zero elements). The Bernoulli distribution parameter

p decides for each element, what state is active and controls the level of sparsity, and then each state is modeled via a Gaussian distribution. It must be taken into account that we only use this procedure for producing the source and assume the user does not have any information about the source probability distribution.

We also need a type of side information about the source that we can model in the form of (3.2). For this purpose we assumed that we have a prior information about the positions and values of some of the large value elements. This assumption is a good embed for testing the proposed method. We transform this information to the form of (3.2). An example makes the procedure clear. Assume we have a sparse source $\mathbf{s} \in \mathbb{R}^{10}$. Assume we know that the second and the fourth elements are non-zero and both equal to 2, thus one concludes:

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{3.15}$$

and:

$$\mathbf{b} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}. \tag{3.16}$$

In the general case for $B \in \mathbb{R}^{m' \times n}$, where m' is the number of the known non-zero elements, in each row we set B_{ij} related to the jth known non-zero element equal to one and the rest of the matrix entries equal to zero. Trivially $\mathbf{b}_j$ is equal to the jth known non-zero element. We continue with experiments in this framework.

3.4.2 Experiments

We set n =1,024, $\sigma_1 = 0.1$, and $\sigma_2 = 10$ with sparsity level of $p = 0.1$ to generate the source signal through this subsection. This means that we have an approximately sparse source with about 100 large value elements. The sensing matrix was chosen as a random matrix with i.i.d Gaussian entries and applied to the source to produce the measurement vector $\mathbf{y}$. This selection is standard in the CS literature because these matrices satisfy both RIP and SSP. Based on our assumption, the positions and the values of a fraction of large value elements are known and one can form (3.2). Through the experiments, we assumed one fourth of the large value elements of the source are known ($m' \approx 25$) and $m = 300$, unless stated.

Figure 3.1a depicts an instance of the generated source in which the signal is presented versus time (index). Figure 3.1b depicts the reconstruction results for the classical CS. Visually, it can be seen some approximately zero elements are estimated larger than the real values and the quality of the reconstruction is poor. Figure 3.1c depicts the reconstruction results for the proposed method. As it can be detected visually our method outperforms the classical method, which is also confirmed numerically by calculating the signal-to-noise ration (SNR). The result confirms that the proposed algorithm works properly and is able to incorporate the side information.

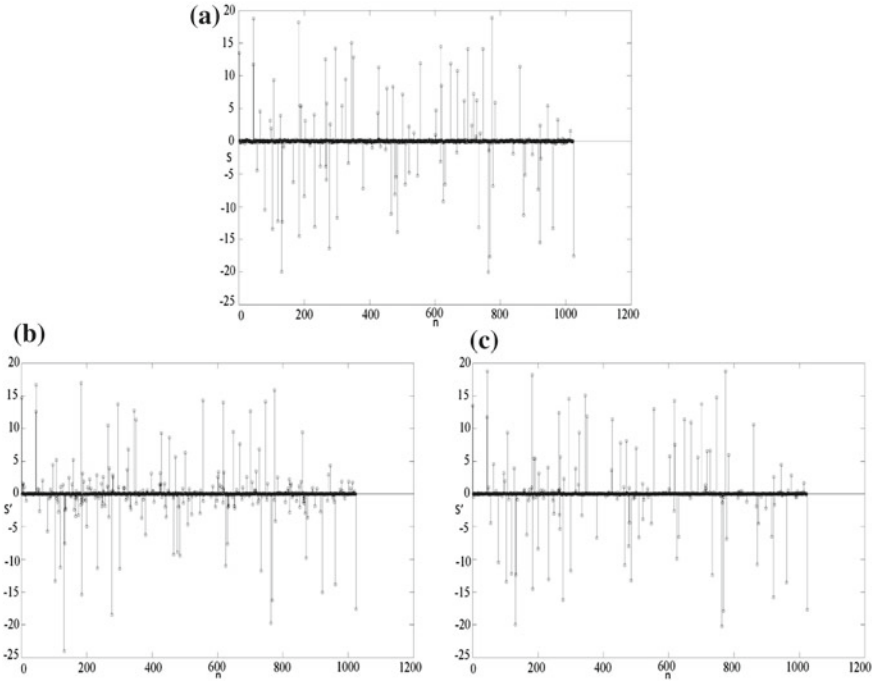

Fig. 3.1 **a** The source signal **b** reconstructed using the classical CS, SNR = 14.4 **c** reconstructed using the proposed method, SNR = 26.0

To analyze the algorithm two sets of experiments are done. First, we study the effect of the number of the known elements on the performance of the algorithm. Assume $0 < r \leq 1$ indicates the fraction of the known elements. Figure 3.2 depicts the proposed algorithm reconstruction quality (measured via SNR) versus r for $r \in (0.1, 1)$. As expected as we increase the side information, the quality of the reconstruction also improves such that for $r = 1$ we have near complete recovery.

In the second experiments we consider the effect of number of the measurements, m, on reconstruction quality. Figure 3.3 depicts output reconstruction SNR versus m for the classical CS and the proposed method. As expected, the reconstruction quality improves as the number of measurement increases for both methods. As it can be detected for large number of the measurements both methods are saturated and we have high SNR values. This is not surprising since when the number of the measurements are large enough we can reconstruct the source perfectly and the side information has negligible effect on the quality of the measurements. But for insufficient measurements, the side information becomes important and improves the quality of the reconstruction. This implies that the proposed method can be used to either improve the quality of the reconstruction or decrease the number of the required measurements without deteriorating the quality of the reconstruction. Overall, these experiments confirm the effectiveness of the proposed method. In the next chapters,

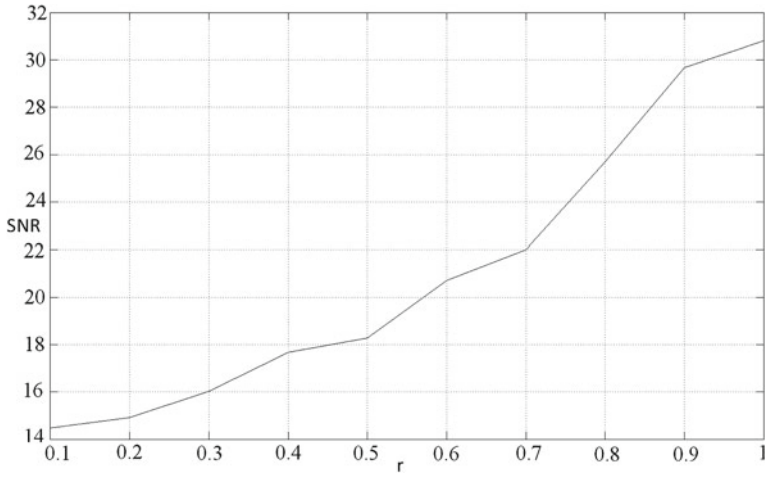

Fig. 3.2 Output SNR versus the fraction of known large value elements (r)

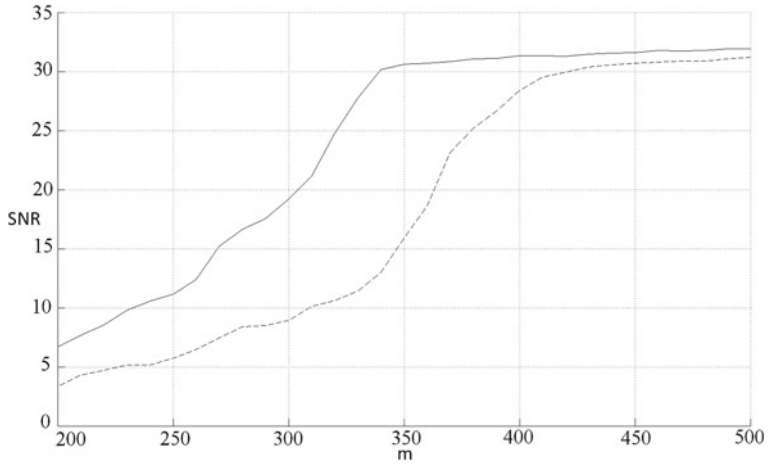

Fig. 3.3 SNR of the source reconstruction obtained with different methods as a function of m. Here, the dashed and solid lines correspond to the classical CS and the proposed CS method, respectively, and $r = 0.25$

this method has been applied to two practical examples: image deblurring in optical imaging [10] and surface reconstruction in the gradient field [11]. In both applications the source signals are gradient fields and the side information can be formulated as (3.2) as in [1]. Also, further analysis has been done through these applications.

References

1. M. Hosseini, O. Michailovich, Derivative compressive sampling with application to phase unwrapping. In *Proceedings of EUSIPCO*, Glasgow, UK, August, 2009
2. S. Osher, M. Burger, D. Goldfarb, J. Xu, W. Yin, An iterative regularization method for total variation-based image restoration. Simul **4**, 460–489 (2005)
3. W. Yin, S. Osher, D. Goldfarb, J. Darbon, Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. SIAM J. Imaging Sci. **1**(1), 143–168 (2008)
4. J. Cai, S. Osher, Z. Shen, Split Bregman methods and frame based image restoration. Multiscale Model. Simul. **8**(2), 337–369 (2009)
5. S.S. Chen, D.L. Donoho, Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
6. I. Daubchies, M. Defrise, C.D. Mol, An iterative thresholding algorithm for linear inverse problems with sparsity constraint. Comm. Pure Appl. Math. **75**, 1412–1457 (2009)
7. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imag. Sci. **2**, 183–202 (2009)
8. Z. Shi, H. Tang, Y. Tang, Blind source separation of more sources than mixtures using sparse mixture models. Pattern Recogn. Lett. **26**(16), 2491–2499 (2005)
9. S.A. Solla, T.K. Leen, K. Müller (eds.), *Advances in Neural Information Processing Systems*, vol. 12 (MIT Press, USA, 2000). [NIPS Conference, Denver, Colorado, USA, November 29–December 4, 1999]
10. M. Rostami, O.V. Michailovich, Z. Wang, Image deblurring using derivative compressed sensing for optical imaging application. IEEE Trans. Image Process. **21**(7), 3139–3149 (2012)
11. M. Rostami, O.V. Michailovich, Z. Wang, Gradient-based surface reconstruction using compressed sensing. In *19th IEEE International Conference on Image Processing*, Oralndo, U.S., 2012