Project Report

Cross Validation in Compressive Sensing

Mohammad Ali Rehan & Shreya Pathak
Under Prof. Ajit Rajwade

1 Introduction

Cross Validation techniques can help improve the performance of greedy algorithms for recovery from sparse measurements. Here we give a theoretical basis and probabilistic performance bounds on OMP-CV and various experimental results.We also look at applications of the same.

2 Theoretical guarantees of OMP CV

Let the signal be represented by x. It has N elements of which at max k are non zero. We make A (the reconstruction matrix) have unit column norm.A has m rows and A_{cv} has m_{cv}. Since the same acquisition system also produces the cross validation matrix, we can in general expect this matrix to have column norm as $\sqrt{m_{cv}/m}$, and can expect every element to contribute $\frac{1}{m}$ towards their column norm. This is because for $m_{cv} = m$ we will get 1 (case of reconstruction matrix).
The error terms we define are

$$\Delta x^p = x - \hat{x^p}$$

$$\epsilon_x^p = ||\Delta x^p||_2^2$$

$$\epsilon_{cv}^p = ||y_{cv} - A_{cv}\hat{x^p}||_2^2$$

where $\hat{x^p}$ is the output of the algorithm in the p^{th} iteration. Ideally we would like the output where ϵ_x^p is minimised. However we do not know x and can thus find only ϵ_{cv}^p and must use this value. y_{cv} is measurement corresponding to A_{cv}, and y corresponding to A.

2.1 Lemmas on RIP

1. By relation between eigenvalues and RIP we know

$$max\left(\frac{||A_T'A_Tx_T||}{||x_T||_2}\right) \leq (1+\delta_k)$$

where $k \geq |T|$. Put $x_T = (A'_T A_T)^{-1} x_T$ to get

$$\frac{||(A'_T A_T)^{-1} x_T||}{||x_T||_2} \leq \frac{1}{(1-\delta_k)}$$

2. Let S and T be two disjoint index sets, used to index columns. Then we have $||A'_S A_T||_2 \leq \delta_{|S|+|T|}$

   Proof- To extract submatrix from a matrix (say B) we can do $C = M_1 B M_2$ where in M_1 and M_2 has ones only on principal diagonals at the indices of the rows and columns we want in C respectively. Since M_1 and M_2 is diagonal, their norm is 1. Now

   $$\|C\| \leq \|M_1\| \|B\| \| |M_2\| \leq \|B\|$$

   Note that $A'_S A_T$ is a submatrix of $A'_R A_R - I$ where R is $S \cup T$. So

   $$||A'_S A_T|| \leq ||A'_R A_R - I||$$

   Maximum eigenvalue of $A'_R A_R - I$ is $1 + \delta_{|R|} - 1$ using the result of RIP and eigenvalues as done in class.

3. $$||A^{\dagger}_S A_T x_T||_2 \leq \frac{\delta_{|S|+|T|} ||x_T||_2}{1-\delta_S}$$

   Proof- replace $A^{\dagger}_S$ in LHS as

   $$||(A'_S A_S)^{-1} A'_S A_T x_T||$$

   Use lemma 1 to get

   $$||(A'_S A_S)^{-1} A'_S A_T x_T|| \leq \frac{||A'_S A_T x_T||}{1-\delta_S}$$

   and then use lemma 2.

4. x and z are unit vectors with disjoint subsets of size $|S|$ and $|T|$ respectively. Now

   $$\langle \Phi x, \Phi z \rangle = \frac{1}{4}(|\Phi x + \Phi z|^2 + |\Phi x - \Phi z|^2)$$

   $x + z$ has sparsity $|S| + |T|$ let the RIC for it by $a = \delta_{|S|+|T|}$. So above equations has

   $$\leq \frac{(1+a)|x+z|^2 - (1-a)|x+z|^2}{4}$$

   $$= \frac{a|x+z|^2}{2}$$

   Now because x and z have disjoint support $|x+z|^2 = 2$ this gives,

   $$\langle \Phi x, \Phi z \rangle \leq \delta_{|S|+|T|}$$

If the norm is not one , then can use $\frac{x}{||x||_2}$ and $\frac{z}{||z||_2}$ in its place which will give,

$$\langle \Phi x, \Phi z \rangle \leq \delta_{|S|+|T|}||x||_2||z||_2 \leq \delta_{|S|+|T|}\frac{||\Phi x||_2||\Phi z||_2}{1-\delta_{|S|+|T|}}$$

We used RIP for $|S|+|T|$ which is allowed since both x and z have sparsity less than or equal to it. Above inequality puts limits on the range of angle between the two arguments of the dot product.

5.

$$A_T x_T = PA_T x_T + (I-P)A_T x_T --(*)$$

The P operator projects a vector onto space spanned by A_S, now clearly

$$||(I-P)A_T x_T|| \leq ||A_T x_T||$$

since the LHS is projection of the vector.
Now we will have a right angle triangle whose hypotenuse is $||A_T x_T||$ and the other two sides are $||PA_T x_T||$ and $||(I-P)A_T x_T||$. Let angle between $||A_T x_T||$ and $||PA_T x_T||$ be θ. So,

$$cos(\theta) = \frac{||PA_T x_T||}{||A_T x_T||} = \frac{\langle A_T x_T, PA_T x_T \rangle}{||PA_T x_T||||A_T x_T||}$$

Since $PA_T x_T$ is projection it can be replaced by $A_S y_S$, since it must be the linear combinations of the s vectors which span this space. We now extend y_S and x_T to size n by replacing all other indices to be 0, and hence replace A_T and A_S by A. Now we can use lemma 4, by putting A in place of Φ, the extended y_S and x_T become the orthogonal x and z (because of disjoint support). Hence we get

$$\frac{||PA_T x_T||}{||A_T x_T||} \leq \frac{\delta_{|S|+|T|}}{1-\delta_{|S|+|T|}}$$

The upper bound on $||PA_T x_T||$ when used in $*$ gives

$$\sqrt{1-\left(\frac{\delta_{|S|+|T|}}{1-\delta_{|S|+|T|}}\right)^2}||A_T x_T||_2 \leq ||(I-P)A_T x_T||_2$$

6. Let $\hat{x}^p$ and $\hat{x}^q$ be the recovered signals in the pth and qth iteration respectively in the OMP-CV algorithm with $p < q$. Let T^i denote the support set of x^i. Let $\Gamma_i = (A'_{T^i}A_{T^i})^{-1}$. Then,

$$A_{T^q} = \begin{bmatrix} A_{T^p} & A_{T^{q-p}} \end{bmatrix}$$

$$A'_{T^q}A_{T^q} = \begin{bmatrix} A'_{T^p} \\ A'_{T^{q-p}} \end{bmatrix} \begin{bmatrix} A_{T^p} & A_{T^{q-p}} \end{bmatrix}$$

$$\Gamma_q = (\begin{bmatrix} A'_{T^p}A_{T^p} & A'_{T^p}A_{T^{q-p}} \\ A'_{T^{q-p}}A_{T^p} & A'_{T^{q-p}}A_{T^{q-p}} \end{bmatrix})^{-1}$$

Define $\alpha_p = A'_{T^p}A_{T^{q-p}}$, $\theta_p = (A'_{T^{q-p}}P_{T^p}A_{T^{q-p}})^{-1} = A^{-1}_{T^{q-p}}P^{-1}_{T^p}(A'_{T^{q-p}})^{-1}$, $P_{T^p} = I - A^{\dagger}_{T^p}$
So, $\Gamma_p\alpha_p\theta_p\alpha'_p\Gamma_p = (A_{T^p})^{-1}P^{-1}_{T^p}(A'_{T^p})^{-1}$, $\Gamma_p\alpha_p\theta_p = (A_{T^p})^{-1}P^{-1}_{T^p}(A_{T^{q-p}})^{-1}$, $\theta_p\alpha'_p\Gamma_p = A^{-1}_{T^{q-p}}P^{-1}_{T^p}(A'_{T^p})^{-1}$
Note: dagger denotes pseudo inverse

$$\begin{bmatrix}\Gamma_p + \Gamma_p\alpha_p\theta_p\alpha'_p\Gamma_p & -\Gamma_p\alpha_p\theta_p\\ -\theta_p\alpha'_p\Gamma_p & \theta_p\end{bmatrix} = \begin{bmatrix}\Gamma_p + (A_{T^p})^{-1}P^{-1}_{T^p}(A'_{T^p})^{-1} & -(A_{T^p})^{-1}P^{-1}_{T^p}(A_{T^{q-p}})^{-1}\\ A^{-1}_{T^{q-p}}P^{-1}_{T^p}(A'_{T^p})^{-1} & (A'_{T^{q-p}}P_{T^p}A_{T^{q-p}})^{-1}\end{bmatrix}$$

$$\begin{bmatrix}\Gamma_p + \Gamma_p\alpha_p\theta_p\alpha'_p\Gamma_p & -\Gamma_p\alpha_p\theta_p\\ -\theta_p\alpha'_p\Gamma_p & \theta_p\end{bmatrix}\begin{bmatrix}A'_{T^p}A_{T^p} & A'_{T^p}A_{T^{q-p}}\\ A'_{T^{q-p}}A_{T^p} & A'_{T^{q-p}}A_{T^{q-p}}\end{bmatrix} = I$$

So, $\Gamma_q = \begin{bmatrix}\Gamma_p + \Gamma_p\alpha_p\theta_p\alpha'_p\Gamma_p & -\Gamma_p\alpha_p\theta_p\\ -\theta_p\alpha'_p\Gamma_p & \theta_p\end{bmatrix}$
Expressing $A^{\dagger}_{T^q}$ in terms of $A^{\dagger}_{T^p}$, we have,

$$A^{\dagger}_{T^q} = (A'_{T^q}A_{T^q})^{-1}A'_{T^q} = \Gamma_q A'_{T^q}$$

From the above derivations, we have,

$$A^{\dagger}_{T^q} = \Gamma_q A'_{T^q} = \begin{bmatrix}\Gamma_p + \Gamma_p\alpha_p\theta_p\alpha'_p\Gamma_p & -\Gamma_p\alpha_p\theta_p\\ -\theta_p\alpha'_p\Gamma_p & \theta_p\end{bmatrix}\begin{bmatrix}A'_{T^p}\\ A'_{T^{q-p}}\end{bmatrix}$$

Using the fact that $\Gamma_p A'_{T^p} = A^{\dagger}_{T^p}$, we get,

$$A^{\dagger}_{T^q} = \begin{bmatrix}A^{\dagger}_{T^p} + A^{\dagger}_{T^p}A_{T^{q-p}}\theta_p A'_{T^{q-p}}A_{T^p}A^{\dagger}_{T^p} - A^{\dagger}_{T^p}A_{T^{q-p}}\theta_p A'_{T^{q-p}}\\ -\theta_p A'_{T^{q-p}}A_{T^p}A^{\dagger}_{T^p} + \theta A'_{T^{q-p}}\end{bmatrix}$$

$$= \begin{bmatrix}A^{\dagger}_{T^p} - A^{\dagger}_{T^p}A_{T^{q-p}}\theta_p A'_{T^{q-p}}P_{T^p}\\ \theta A'_{T^{q-p}}P_{T^p}\end{bmatrix}$$

Now, $\hat{x}^q = A^{\dagger}_{T^q}y = A^{\dagger}_{T^q}(Ax+n) = A^{\dagger}_{T^q}(A_{T^q}x_{T^q} + A_{(T^q)^c}x_{(T^q)^c} + n) = A^{\dagger}_{T^q}(A_{(T^q)^c}x_{(T^q)^c} + n) + x_{T^q} =$

$$\begin{bmatrix}A^{\dagger}_{T^p} - A^{\dagger}_{T^p}A_{T^{q-p}}\theta_p A'_{T^{q-p}}P_{T^p}\\ \theta_p A'_{T^{q-p}}P_{T^p}\end{bmatrix}(A_{(T^q)^c}x_{(T^q)^c} + n) + x_{T^q}$$

Define $\theta_p A'_{T^{q-p}}P_{T^p}(A_{(T^q)^c}x_{(T^q)^c} + n) = \delta_{T^{q-p}}$.

$$\hat{x}^q = \begin{bmatrix}A^{\dagger}_{T^p}((A_{(T^q)^c}x_{(T^q)^c} + n) - A_{T^{q-p}}\delta_{T^{q-p}})\\ \delta_{T^{q-p}}\end{bmatrix} + x_{T^q}$$

7. We try to estimate the absolute value of the $\delta_{T^{q-p}} = (A'_{T^{q-p}}P_{T^p}A_{T^{q-p}})^{-1}A'_{T^{q-p}}P_{T^p}$ $(A_{(T^q)^c}x_{(T^q)^c} + n)$ vector defined in the previous lemma. Let u be a R^{q-p} dimensional vector. Since the singular values of a matrix and its transpose are same by SVD, we have,

$$\sqrt{1-\delta_{q-p}}||P_{T^p}A_{T^{q-p}}u||_2 \leq ||A'_{T^{q-p}}P_{T^p}A_{T^{q-p}}u||_2 \leq \sqrt{1+\delta_{q-p}}||P_{T^p}A_{T^{q-p}}u||_2$$

Using Lemma 5 defined in this section, we have,

$$\sqrt{1-(\frac{\delta_q}{1-\delta_q})^2}\sqrt{1-\delta_{q-p}}||A_{T^{q-p}}u||_2 \leq ||A'_{T^{q-p}}P_{T^p}A_{T^{q-p}}u||_2 \leq \sqrt{1+\delta_{q-p}}||A_{T^{q-p}}u||_2$$

Using Lemma 1, we get,

$$\sqrt{1-(\frac{\delta_q}{1-\delta_q})^2}(1-\delta_{q-p})||u||_2 \leq ||A'_{T^{q-p}}P_{T^p}A_{T^{q-p}}u||_2 \leq (1+\delta_{q-p})||u||_2$$

By relation between eigenvalues and RIP, we get that all the singular values of $A'_{T^{q-p}}P_{T^p}A_{T^{q-p}}$ lie between $\sqrt{1-(\frac{\delta_q}{1-\delta_q})^2}(1-\delta_{q-p})$ and $(1+\delta_{q-p})$. Thus, the singular values of the inverse of this matrix lie between the inverses of the 2 bounds by SVD. So, we get,

$$\frac{1}{(1+\delta_{q-p})}||u||_2 \leq ||(A'_{T^{q-p}}P_{T^p}A_{T^{q-p}})^{-1}u||_2 \leq \frac{1}{\sqrt{1-(\frac{\delta_q}{1-\delta_q})^2}(1-\delta_{q-p})}||u||_2$$

$$\begin{aligned}||\delta_{T^{q-p}}||_2^2 &= ||(A'_{T^{q-p}}P_{T^p}A_{T^{q-p}})^{-1}A'_{T^{q-p}}P_{T^p}(A_{(T^q)^c}x_{(T^q)^c}+n)||_2^2 \\ &\leq \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}||A'_{T^{q-p}}P_{T^p}(A_{(T^q)^c}x_{(T^q)^c}+n)||_2^2\end{aligned}$$

Expanding $P_{T^p} = I - A_{T^p}A^{\dagger}_{T^p}$, we get,

$$= \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}||A'_{T^{q-p}}(A_{(T^q)^c}x_{(T^q)^c}+n) - A'_{T^{q-p}}A_{T^p}A^{\dagger}_{T^p}(A_{(T^q)^c}x_{(T^q)^c}+n)||_2^2$$

Applying Cauchy-Schwarz inequality we have,

$$\leq \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}(||A'_{T^{q-p}}(A_{(T^q)^c}x_{(T^q)^c}+n)||_2^2 + ||A'_{T^{q-p}}A_{T^p}A^{\dagger}_{T^p}(A_{(T^q)^c}x_{(T^q)^c}+n)||_2^2)$$

Representing $A_{(T^q)^c}x_{(T^q)^c} + n = A_g x_g$, and applying lemma 2 we have,

$$\leq \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}((\delta_{|(T^q)^c|+q-p+1})^2||x_g||_2^2 + \delta_q^2||A^{\dagger}_{T^p}A_g x_g||_2^2)$$

Applying Lemma 3 we get,

$$\leq \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}((\delta_{|(T^q)^c|+q-p+1})^2||x_g||_2^2 + \delta_q^2(\frac{\delta_{p+|(T^q)^c|+1}}{1-\delta_p})^2||x_g||_2^2)$$

Finally, replacing $||x_g||_2^2 = ||[x'_{(T^q)^c}, \sigma_n]||_2^2 = ||x_{(T^q)^c}||_2^2 + \sigma_n^2$, we have,

$$= \frac{1}{1-(\frac{\delta_q}{1-\delta_q})^2(1-\delta_{q-p})^2}((\delta_{|(T^q)^c|+q-p+1})^2 + \delta_q^2(\frac{\delta_{p+|(T^q)^c|+1}}{1-\delta_p})^2)(||x_{(T^q)^c}||_2^2 + \sigma_n^2)$$

$$= \eta(||x_{(T^q)^c}||_2^2 + \sigma_n^2)$$

If $\delta_d < 0.1$ since all footnotes of RIC do not exceed d, we have $\eta \leq 0.0127$

2.2 Bounds on value of Recovery error

We first talk about the cross validation error $\epsilon_{cv} = ||y_{cv} - A_{cv}\hat{x}||^2$ wherein $\hat{x}$ is some recovered value of estimate of the original signal. Put $y_{cv} = A_{cv}x + n_{cv}$ and let $x - \hat{x} = t$ to get

$$\epsilon_{cv} = ||A_{cv}t + n_cv||^2$$

$$\epsilon_{cv} = \sum_{i=0}^{m_{cv}} r_i^2$$

where $r_i = (\sum_{j=0}^{N} a_{ij}t_j) + n_{cvi}$
a_{ij} is an element of A_{cv}. Clearly r_i is Gaussian since its sum of Gaussian and all r_i's are independent. Now $E(r_i) = 0$ since $E(a_{ij}) = 0$ and

$$Var(r_i) = (\sum_{j=0}^{N} \frac{t_j^2}{m}) + \frac{\sigma_n^2}{m} = \frac{||t||^2}{m} + \frac{\sigma_n^2}{m}$$

We assume noise is Gaussian, and σ_n^2 is just a variance parameter. Thus we can write

$$r_i^2 = (\frac{||t||^2}{m} + \frac{\sigma_n^2}{m})\chi_1^2 = b\chi_1^2$$

Clearly the RHS is independent of i , so all r_i's have the same distribution(chi square of degree of freedom 1 and a constant multiplied). We can talk about their sum by using CLT. Hence if their are enough measurements the distribution of the sum (ϵ_{cv}) is Gaussian, with

$$\mu = \frac{m_{cv}}{m}(||t||^2 + \sigma_n^2)$$

and

$$Var = \sigma^2 = \frac{2m_{cv}}{m}(||t||^2 + \sigma_n^2)^2$$

This variance we found based on the variance formula of chi square random variable of 1 degree of freedom which we know is 2.

A basic consequence of the above result is the following theorem (Theorem 1) which relates the cv residual ϵ_{cv} and the recovery error $\epsilon_x = ||t||^2$ described above.
For a standard normal distribution we know that $\Phi(n) - \Phi(-n) = erf(n/\sqrt{2})$ where Φ is the CDF of the standard Gaussian and $erf(.)$ is the error function. Hence for a Gaussian $x \sim N(\mu, \sigma^2)$, the following bound holds with probability $erf(n/\sqrt{2})$

$$\mu - n\sigma \leq x \leq \mu + n\sigma$$

Applying this to ϵ_{cv}, the following inequality holds with probability $erf(\lambda/\sqrt{2})$

$$\frac{m_{cv}}{m}(\epsilon_x + \sigma_n^2) - \lambda\frac{\sqrt{2m_{cv}}}{m}(\epsilon_x + \sigma_n^2) \leq \epsilon_{cv} \leq \frac{m_{cv}}{m}(\epsilon_x + \sigma_n^2) + \lambda\frac{\sqrt{2m_{cv}}}{m}(\epsilon_x + \sigma_n^2)$$

This can be simplified to

$$\frac{m_{cv}}{m}(1-\lambda\sqrt{\frac{2}{m_{cv}}})(\epsilon_x+\sigma_n^2) \leq \epsilon_{cv} \leq \frac{m_{cv}}{m}(1+\lambda\sqrt{\frac{2}{m_{cv}}})(\epsilon_x+\sigma_n^2)$$

Defining $h(\lambda,\pm) = \frac{m}{m_{cv}}\frac{1}{1\pm\lambda\sqrt{2/m_{cv}}}$ and rearranging we get,

$$\frac{\epsilon_{cv}}{h(\lambda,+)} \leq \epsilon_x+\sigma_n^2 \leq \frac{\epsilon_{cv}}{h(\lambda,-)}$$

$$\frac{\epsilon_{cv}}{h(\lambda,+)} - \sigma_n^2 \leq \epsilon_x \leq \frac{\epsilon_{cv}}{h(\lambda,-)} - \sigma_n^2 \tag{1}$$

This gives an estimate of the interval of the interval of the recovery error by the observed CV residual with probability $erf(\lambda/\sqrt{2})$, i.e., $\frac{m}{m_{cv}}\frac{2\lambda\sqrt{2}}{\sqrt{m_{cv}-2\lambda^2/\sqrt{m_{cv}}}}\epsilon_{cv}$. Thus the interval decreases in size as number of CV measurements increases as expected.

2.3 Comparison of Recovery error

We now compare 2 signals $\hat{x}^p$ and $\hat{x}^q$. Theorem 1 gives a relation between the recovery error and the CV residual. We now formulate this.

For the purpose of simplicity we define $A_g = [A, a_g]$ and $x_g = [x'\sigma_n]'$, then $y = Ax + n = A_g x_g$. Note that the noise cannot be recovered thus we extend our original definitions to get $\Delta x_g^p = [(\Delta x^p)', \sigma_n]'$ and $\epsilon_g^p = ||\Delta x_g^p||_2^2$. Also define $\Delta\epsilon_{cv} = \epsilon_{cv}^p - \epsilon_{cv}^q$.

Let $x - \hat{x}^p = t$, then we can write, $\Delta\epsilon_{cv} = \sum_{i=1}^{m_{cv}} r_i$ where $r_i = ((\sum_{j=0}^N a_{ij}t_j) + n_{cvi})^2 - ((\sum_{k=0}^N a_{ik}t_j) + n_{cvi})^2$ or

$$r_i = (\sum_{j=1}^N a_{cv_{ij}}(t_j^p + t_j^q))(\sum_{k=1}^N a_{cv_{ik}}(t_k^p - t_k^q)) + 2n_{cv_i}(\sum_{k=1}^N a_{cv_{ik}}(t_k^p - t_k^q))$$

Expanding r_i we observe that it is a linear combination of terms of the form $a_{cv_{ij}}^2$, $a_{cv_{ij}}a_{cv_{ik}}$ and $n_{cv_i}a_{cv_{ik}}$. Also, using the formulae of mean and variance of product of independent random variables we have, $E(a_{cv_{ij}}^2) = 1/m$, $E(a_{cv_{ij}}a_{cv_{ik}}) = 0$, $E(n_{cv_i}a_{cv_{ik}}) = 0$ and $Var(a_{cv_{ij}}^2) = 2/m^2$, $Var(a_{cv_{ij}}a_{cv_{ik}}) = 1/m^2$, $Var(n_{cv_i}a_{cv_{ik}}) = \sigma_n^2/m^2$.

Using the fact that each of these terms are mutually independent, we get

$$E(r_i) = \frac{1}{m}(\sum_{j=1}^N (t_j^p + t_j^q)(t_j^p - t_j^q)) = \frac{1}{m}((t_j^p)^2 - (t_j^q)^2) = \frac{1}{m}(\epsilon_g^p - \epsilon_g^q)$$

$$Var(r_i) = Var(\sum_{j=1}^N a_{cv_{ij}}^2(t_j^p + t_j^q)(t_j^p - t_j^q)) +$$

$$Var(\sum_{j=1}^N \sum_{k=j+1}^N a_{cv_{ij}}a_{cv_{ik}}((t_j^p + t_j^q)(t_k^p - t_k^q) + (t_k^p + t_k^q)(t_j^p - t_j^q)) +$$

$$4Var(\sum_{j=1}^{N} n_{cv_i} a_{cv_{ij}} (t_j^p - t_j^q))$$

For the variance, the first two terms can be written as

$$2/m^2(\sum_{j=1}^{N}((t_j^p)^2 - (t_j^q)^2)^2) + 4/m^2(\sum_{j=1}^{N}\sum_{k=j+1}^{N}(t_j^p t_k^p - t_j^q t_k^q)^2) = 2/m^2((\sum (t_j^p)^4$$

$$+2\sum_{j=1}^{N}\sum_{k=j+1}^{N}(t_j^p)^2(t_k^p)^2)+(\sum (t_j^q)^4+2\sum_{j=1}^{N}\sum_{k=j+1}^{N}(t_j^q)^2(t_k^q)^2)-2(\sum (t_j^p)^2(t_j^q)^2+2\sum_{j=1}^{N}\sum_{k=j+1}^{N} t_j^p t_j^q t_k^p t_k^q))$$

$$= 2/m^2((\sum_{j=1}^{N}(t_j^p)^2)^2 + (\sum_{j=1}^{N}(t_j^q)^2)^2 - 2(\sum_{j=1}^{N} t_j^p t_j^q)^2) = 2/m^2((\epsilon_x^p)^2 + (\epsilon_x^q)^2 - 2\langle t^p, t^q \rangle^2)$$

The third term can be simplified to $4\sigma_n^2/m^2 \sum_{j=1}^{N}(t_j^p - t_j^q)^2 = 4\sigma_n^2/m^2 ||t^p - t^q||^2$

$$Var(r_i) = \frac{2}{m^2}((\epsilon_x^p)^2 + (\epsilon_x^q)^2 - 2\langle t^p, t^q \rangle^2 - 2\sigma_n^2 ||t^p - t^q||^2)$$

Defining $\rho_g = \frac{\langle \Delta x_g^p, \Delta x_g^q \rangle}{||\Delta x_g^p||_2 ||\Delta x_q^q||_2}$ and using the fact that $\epsilon_g^p = \epsilon_x^p + \sigma_n^2$ and $\langle \Delta x_g^p, \Delta x_g^q \rangle = \langle \Delta x^p, \Delta x^q \rangle + \sigma_n^2$, we simplify this to

$$Var(r_i) = \frac{2}{m^2}((\epsilon_g^p)^2 + (\epsilon_g^q)^2 - 2\rho_g \epsilon_g^p \epsilon_g^q$$

Since, m_{cv} is large enough, we can apply CLT to get

$$\Delta\epsilon_{cv} = \sum_{i=1}^{m_{cv}} r_i \sim N(\mu, \sigma^2)$$

where $\mu = \frac{m_{cv}}{m}(\epsilon_g^p - \epsilon_g^q)$ and $\sigma^2 = \frac{2m_{cv}}{m^2}((\epsilon_g^p)^2 + (\epsilon_g^q)^2 - 2\rho_g \epsilon_g^p \epsilon_g^q)$

We see that if $\epsilon^p \geq \epsilon^q$ then $\epsilon_g^p \geq \epsilon_g^q$ Thus the μ parameter for the Random variable $\Delta\epsilon_{cv}$ is positive. We now want the probability that $\Delta\epsilon_{cv}$ is positive (Theorem2). This is the probability that a standard normal variable is greater than $-k = -\frac{\mu}{\sigma}$. This probability is

$$\Phi(k) = \frac{1}{\sqrt{2\pi}}\int_{-k}^{\infty} e^{\frac{-t^2}{2}} dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{k} e^{\frac{-t^2}{2}} dt$$

by putting t as $-t$. Now,

$$\frac{1}{k^2} = \frac{2}{m_{cv}}\left[1 + 2(1 - \rho_g^2)\frac{\epsilon_g^p \epsilon_g^q}{(\epsilon_g^p - \epsilon_g^q)^2}\right]$$

This probability can increase by larger m_{cv} meaning more measurements. Also a larger correlation (and hence similarity) between two measurements being compared leads to the same ordering between the two error terms. The effect of the value of the error terms is studied below.

Let k also be denoted by λ. Suppose we want $\lambda > \lambda_0$ then, assuming $\frac{\epsilon_g^p}{\epsilon_g^q}$ as a single variable we get the relation ,

$$\frac{\epsilon_g^p}{\epsilon_g^q} + \frac{\epsilon_g^q}{\epsilon_g^p} \geq 2C + 2$$

where $C = \frac{\lambda_0^2(1-\rho_g^2)}{m_{cv}-2\lambda_0^2}$

By using the quadratic formula and the condition that $\frac{\epsilon_g^p}{\epsilon_g^q} \geq 0$ we get that

$$\frac{\epsilon_g^p}{\epsilon_g^q} \geq 2C + 1 + 2\sqrt{C^2 + C}$$

Thus more the difference in magnitudes of recovery error, larger is the probability that CV error follows the same ordering.

2.4 Noise Robust OMP

Here we place probabilistic bounds on the condition that the iterations which recover incomplete support set have larger CV error than that of oracle output with high probability.

Also the iterations which do recover the complete support set and is also the output of OMP-CV has a recovery error within a constant factor of the oracle output's recovery error.

We start with lemma 6 of section 2.1 by replacing q with p. In this case $\delta_{T^{p-p}}$ is the zero vector since $A'_{T^{p-p}}$ is zero matrix.So we get

$$x_{T^p} - \hat{x}^p_{T^p} = -(A_{T^p})^\dagger (A_{(T^p)^c} x_{(T^p)^c} + n)$$

Note that x_{T^p} is the sub-vector of original signal indexed by the support set of the p^{th} iteration. $\hat{x}^p_{T^p}$ is the sub vector of the recoverd vector after p iterations indexed by the support set T^p

$$\Delta x^p = \begin{bmatrix} x_{T^p} - \hat{x}^p_{T^p} \\ x_{T^{q-p}} \\ x_{(T^q)^c} \end{bmatrix} = \begin{bmatrix} -(A_{T^p})^\dagger (A_{(T^p)^c} x_{(T^p)^c} + n) \\ x_{T^{q-p}} \\ x_{(T^q)^c} \end{bmatrix}$$

where $q > p$ is a later iteration having a support set which is the super set of T^p.We order the vector Δx^p in the order the indices get included in the support set.

Using the same lemma and similar ideas, we can write

$$\Delta x^q = \begin{bmatrix} x_{T^q} - \hat{x}^q_{T^q} \\ x_{(T^q)^c} \end{bmatrix} = \begin{bmatrix} -(A_{T^p})^\dagger (A_{(T^q)^c} x_{(T^q)^c} + n - A_{T^{q-p}} \delta_{T^{q-p}}) \\ -\delta_{T^{q-p}} \\ x_{(T^q)^c} \end{bmatrix}$$

where $T^{q-p} = T^q \backslash T^p$
Let

$$a = -(A_{T^p})^{\dagger}(A_{(T^p)^c}x_{(T^p)^c} + n) = [(A_{g_{T^p}})^{\dagger}](A_{g_R}x^p_{g_R})$$

R is the set of $(T^p)^c$ along with the last column to include the noise term. We have replaced above using the generalised terms described in section 2.3. Now applying lemma 3 to get

$$||a||_2^2 \leq \left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2 (||x_{(T^p)^c}||_2^2 + \sigma_n^2)$$

Similarly define b as the negative of the first of the three component vector in Δx^q above. In generalised terms,

$$b = [(A_{g_{T^p}})^{\dagger}](A_{g_Q}x_{g_Q} - A_{T^{q-p}}\delta_{T^{q-p}})$$

wherein Q is $(T^q)^c$ along with last column. Note that $|T^p| + |Q| \leq k+1$ maximum sparsity of generalised vector and $|T^p| + |T^{q-p}| \leq k+1$ and both pairs are disjoint as well. We use lemma 3 again but replacing the δ's in numerator with δ_{k+1} since it's larger than both. This gives

$$||b||_2^2 \leq \left(\frac{\delta_{k+1}}{1-\delta_p}\right)^2 (||\delta_{T^{q-p}}||_2^2 + ||x_{(T^q)^c}||_2^2 + \sigma_n^2)$$

Since T^p is a subset of T^q in b we can do the following replacement in b,

$$A_{(T^q)^c}x_{(T^q)^c} = A_{(T^p)^c}x_{(T^p)^c} - A_{T^{q-p}}x_{T^{q-p}}$$

With this replacement

$$c = a - b = (A_{T^p})^{\dagger}(A_{T^{q-p}}(x_{T^{q-p}} + \delta_{T^{q-p}}))$$

Now again by lemma 3,

$$||c||_2^2 \leq \left(\frac{\delta_q}{1-\delta_p}\right)^2 ||x_{T^{q-p}} + \delta_{T^{q-p}}||_2^2$$

Above we derived in theorem 2, the parameter k for the probability that $\Delta\epsilon_{cv} \geq 0$, we study it here further by replacing ρ_g in terms of Δx_g. So, the term in k

$$(1-\rho_g^2)\frac{\epsilon_g^p\epsilon_g^q}{(\epsilon_g^p - \epsilon_g^q)^2} = \frac{\epsilon_g^p\epsilon_g^q - \langle\Delta x_g^p, \Delta x_g^q\rangle^2}{(\epsilon_g^p - \epsilon_g^q)^2} - -(**)$$

Now

$$\epsilon_g^p = \sigma_n^2 + ||x_{T^{q-p}}||_2^2 + ||x_{(T^q)^c}||_2^2 + ||a||_2^2$$

$$\epsilon_g^q = \sigma_n^2 + ||\delta_{T^{q-p}}||_2^2 + ||x_{(T^q)^c}||_2^2 + ||b||_2^2$$

$$\langle\Delta x_g^p, \Delta x_g^q\rangle = a.b - x_{T^{q-p}}.\delta_{T^{q-p}} + ||x_{(T^q)^c}||_2^2 + \sigma_n^2$$

We study the numerator in $(**)$ by substituting the above three equations and when opening brackets we keep $||x_{(T^q)^c}||_2^2 + \sigma_n^2$ together. This gives

$$\epsilon_g^p \epsilon_g^q - \langle \Delta x_g^p, \Delta x_g^q \rangle^2 = (||x_{T^{q-p}}||_2^2 + ||a||_2^2)(||\delta_{T^{q-p}}||_2^2 + ||b||_2^2) - (a.b - x_{T^{q-p}}.\delta_{T^{q-p}})^2$$

$$+(||x_{(T^q)^c}||_2^2 + \sigma_n^2)(||x_{T^{q-p}} + \delta_{T^{q-p}}||_2^2 + ||a-b||_2^2)$$

Now ignore the term being subtracted and use the upper bounds on a, b and $c = a - b$ to get

$$\epsilon_g^p \epsilon_g^q - \langle \Delta x_g^p, \Delta x_g^q \rangle^2 \leq \left[||x_{T^{q-p}}||_2^2 + \left(\frac{\delta_{k+1}}{1-\delta_p} \right)^2 (||x_{(T^p)^c}||_2^2 + \sigma_n^2) \right]$$

$$\left[||\delta_{T^{q-p}}||_2^2 + \left(\frac{\delta_{k+1}}{1-\delta_p} \right)^2 (||x_{(T^q)^c}||_2^2 + \sigma_n^2) \right]$$

$$\left(1 + \left(\frac{\delta_q}{1-\delta_p} \right)^2 \right) ||x_{T^{q-p}} + \delta_{T^{q-p}}||_2^2 (||x_{(T^q)^c}||_2^2 + \sigma_n^2) - (***)$$

When $q = o$, i.e. the oracle output iteration($T \subset T^o$) then $||x_{(T^o)^c}||_2 = 0$ because $(T^o)^c = T \backslash T^o = \phi$, also $||x_{(T^p)^c}||_2 \leq ||x_{T^{o-p}}||_2$ since $x_{(T^p)^c}$ has lesser elements. Use $q = o$ and $||x_{(T^o)^c}||_2 = 0$ in lemma 7 to get $||\delta_{T^{o-p}}||_2^2 \leq \eta \sigma_n^2$
Plugging this in $(***)$ gives

$$\epsilon_g^p \epsilon_g^q - \langle \Delta x_g^p, \Delta x_g^q \rangle^2 \leq \frac{\beta_1}{2} (\alpha^p)^2 \sigma_n^4 + \frac{\beta_2}{2} \sigma_n^4$$

where β_1 and β_2 depend only on RIC and η . Similarly the denominator in $(**)$ for $q = o$ is lower bounded as below:

$$\epsilon_g^p - \epsilon_q^o = ||x_{T^{o-p}}||_2^2 - ||\delta_{T^{o-p}}||_2^2 + ||a||_2^2 - ||b||_2^2$$

Use $a = b + c$ so $||a||_2^2 - ||b||_2^2 = ||c||_2^2 + 2\langle b, c \rangle$
b with $q = o$ (use that x_{T^oC} has no elements)

$$b = (A_{T^p})^\dagger (n - A_{T^{o-p}} \delta_{T^{o-p}})$$

Putting this in the above formula to find $||a||_2^2 - ||b||_2^2$ gives

$$\epsilon_g^p - \epsilon_q^o = ||x_{T^{o-p}}||_2^2 + ||[(A_{T^p})^\dagger] A_{T^{o-p}} x_{T^{o-p}}||_2^2 - ||\delta_{T^{o-p}}||_2^2 -$$

$$||(A_{T^p})^\dagger A_{T^{o-p}} \delta_{T^{o-p}}||_2^2 + 2\langle (A_{T^p})^\dagger A_{T^{o-p}} (\delta_{T^{o-p}} + x_{T^{o-p}}), (A_{T^p})^\dagger n \rangle$$

Now we make this quantity as small as possible. For this we ignore the second term. Replace $||\delta_{T^{o-p}}||_2^2$ using lemma 7. For the fourth term use lemma 3. For the dot product use Cauchy Schwarz and then lemma 3 for the term corresponding to first argument of the dot product. Then use triangular inequality and lemma 7 after it. For the second term we want to maximise its norm, so we make a generalised term out of it. We have only the error term , for the rest we put zeros, since lemma 3 requires the two matrices who have disjoint support. The largest such set possible is $(T^p)^c$. After all this, substitute $||x_{T^{o-p}}|| \geq \alpha^p\sigma_n$ because the sub vector of x in RHS has smaller index set. All this is shown below.

$$\begin{aligned}
\varepsilon_{\mathrm{g}}^{p} - \varepsilon_{\mathrm{g}}^{o} &= \|\mathbf{x}_{T^{o-p}}\|_2^2 - \|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 + \|\mathbf{a}\|_2^2 - \|\mathbf{b}\|_2^2 \\
&= (\|\mathbf{x}_{T^{o-p}}\|_2^2 + \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\mathbf{x}_{T^{o-p}}\|_2^2) - (\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 + \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2^2) \\
&\quad + 2\langle \mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}(\mathbf{x}_{T^{o-p}} + \boldsymbol{\delta}_{T^{o-p}}), \mathbf{A}_{T^p}^{\dagger}\mathbf{n}\rangle \\
&\geq \|\mathbf{x}_{T^{o-p}}\|_2^2 - (1 + (\frac{\delta_o}{1-\delta_p})^2)\|\boldsymbol{\delta}_{T^{o-p}}\|_2^2 - 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}\|\mathbf{x}_{T^{o-p}} + \boldsymbol{\delta}_{T^{o-p}}\|_2\sigma_{\mathrm{n}} \qquad (82) \\
&\geq (\alpha^p)^2\sigma_{\mathrm{n}}^2 - (1 + (\frac{\delta_o}{1-\delta_p})^2)\eta\sigma_{\mathrm{n}}^2 - 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}((\alpha^p)^2\sigma_{\mathrm{n}} + \sqrt{\eta}\sigma_{\mathrm{n}})\sigma_{\mathrm{n}} \\
&\geq \left[(\alpha^p)^2 - \beta_3\alpha^p - \beta_4\right]\sigma_{\mathrm{n}}^2,
\end{aligned}$$

where $\beta_3 = 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}$ and $\beta_4 = (1 + (\frac{\delta_o}{1-\delta_p})^2)\eta + 2\frac{\delta_o\delta_{p+1}}{(1-\delta_p)^2}\sqrt{\eta}$. Notice that $\varepsilon_{\mathrm{g}}^{p} - \varepsilon_{\mathrm{g}}^{o} > 0$, then we have,

$$\varepsilon_{\mathrm{g}}^{p} - \varepsilon_{\mathrm{g}}^{o} \geq \max(\left[(\alpha^p)^2 - \beta_3\alpha^p - \beta_4\right]\sigma_{\mathrm{n}}^2, 0). \qquad (83)$$

This helps us give the relation

$$\frac{1}{\lambda^2} \leq \frac{2}{m_{cv}}\left(1 + \frac{\beta_1(\alpha^p)^2 + \beta_2}{(max((\alpha^p)^2 - \beta_3\alpha^p - \beta_4, 0)^2}\right) - -(\#)$$

This gives

$$\lambda^2 \geq \frac{m_{cv}}{2}(1 - g(\alpha^p))$$

where $\mathrm{g}(\alpha^{\mathrm{p}}) = [\beta_1(\alpha^{\mathrm{p}})^2 + \beta_2]/[\beta_1(\alpha^{\mathrm{p}})^2 + \beta_2 + \max((\alpha^{\mathrm{p}})^2 - \beta_3\alpha^{\mathrm{p}} - \beta_4, 0)^2]$.

Since β_1 is mostly much larger than all others we get

$$\lambda^2 \geq \frac{m_{cv}}{2}\left(1 - \frac{\beta_1}{(\alpha^p)^2 + \beta_1}\right)$$

We consider the other case now, when $T \backslash T^p = \phi$ As we have done above

$$\Delta x^p = \begin{bmatrix} -(A_{T^p})^{\dagger}(A_{(T^p)^c}x_{(T^p)^c} + n) \\ x_{T^{o-p}} \\ x_{(T^o)^c} \end{bmatrix}$$

Note that $x_{T^{o-p}}$, $x_{(T^p)^c}$ and $x_{(T^o)^c}$ are zero vectors since they don't include any index in T. So,

$$
\begin{aligned}
\langle \Delta \mathbf{x}_{\mathrm{g}}^{o}, \Delta \mathbf{x}_{\mathrm{g}}^{p} \rangle &= \langle \mathbf{A}_{T^p}^{\dagger}\mathbf{n}, \mathbf{A}_{T^p}^{\dagger}(\mathbf{n} - \mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}) \rangle + \sigma_{\mathrm{n}}^2 \\
&= \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2^2 - \langle \mathbf{A}_{T^p}^{\dagger}\mathbf{n}, \mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}} \rangle + \sigma_{\mathrm{n}}^2 \\
&\geq \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2^2 - \|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2 \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2 + \sigma_{\mathrm{n}}^2 \\
&\geq -\|\mathbf{A}_{T^p}^{\dagger}\mathbf{n}\|_2 \|\mathbf{A}_{T^p}^{\dagger}\mathbf{A}_{T^{o-p}}\boldsymbol{\delta}_{T^{o-p}}\|_2 + \sigma_{\mathrm{n}}^2 \\
&\geq \sigma_{\mathrm{n}}^2 (1 - \frac{\delta_{p+1}\delta_{o+1}}{(1-\delta_p)^2}\sqrt{\eta}),
\end{aligned}
$$

The dagger means pseudo inverse. In the third step we used Cauchy Schwarz and then lemma 3 by converting n to a generalised term with support as $(T^p)^c$, and then lemma 7 on δ term. We now upper bound

$$\epsilon_g^p \epsilon_g^o = (\|(A_{T^p})^{\dagger} n\|_2^2 + \sigma_n^2)(\|(A_{T^p})^{\dagger}(n - A_{T^{o-p}}\delta_{T^{o-p}})\|_2^2 + \|\delta_{T^{o-p}}\|_2^2 + \sigma_n^2)$$

Now we bound each term as follows. The first term by making a generalised vector with all zeros(but with support set $(T^p)^c$)) and use lemma 3 to get

$$\|(A_{T^p})^{\dagger} n\|_2^2 \leq \left(\frac{\delta_{p+1}}{1-\delta_p}\right)^2 \sigma_n^2$$

For the first term in second bracket make a generalised term as $[\delta_{T^{o-p}} \sigma_n]$ and corresponding $A_{gT^{o-p}} = [A_{T^{o-p}} a]$. Now use lemma 3, followed by triangular inequality on $[\delta_{T^{o-p}} \sigma_n]$ and lemma 7 to get

$$\epsilon_g^p \epsilon_g^o \leq \gamma \sigma_n^4$$

after taking σ_n common, γ depends only on RICs and η. This helps us to give

$$\rho_g = \frac{\langle \Delta x_g^p, \Delta x_g^q \rangle}{\sqrt{\epsilon_g^o \epsilon_g^p}} \geq \beta_5$$

β_5 depends only on RICs and η.
if $o > p$ just switch o and p in above argument.
Toward end of section 3 we have that $C = \frac{\lambda_0^2(1-\rho_g^2)}{m_{cv} - 2\lambda_0^2}$
and we get that

$$\frac{\epsilon_g^p}{\epsilon_g^q} \geq 2C + 1 + 2\sqrt{C^2 + C} = C_1$$

now C can be replaced by a constant by replacing ρ_g by β_5 which has no dependence on signal values.
So with probability $\Phi(k_0)$ (which is ensured by $\epsilon_g^p > C_1 \epsilon_g^o$. Also, $\epsilon_g^p > \epsilon_g^o$ as the RHS is oracle output) we have $\epsilon_{cv}^p > \epsilon_{cv}^q$ Consider the set of such indices ($> o$) which have the above probability let their be n elements.
If the output of OMP-CV belongs to this set, then there must be at least 1 element(i)

in the set such that such that $\epsilon_{cv}^{i} \leq \epsilon_{cv}^{o}$. The probability of this happening is $< n(1 - \Phi(k_0)) < (d-k)(1-\Phi(k_0))$
as max value of n is $d-k$, at iteration k we cover entire set T and d are max number of iterations.
The probability of this not happening, that is no iteration that has $\epsilon_g^p > C_1 \epsilon_g^o$ can be output of OMP-CV is $> 1-(d-k)(1-\Phi(k_0))$. This proves theorem 4.
The probability that oracle output covers the entire support set is large, suppose that it didn't. Then any iteration that occurs after it would have covered more elements. So now the cross validation error is expected to be smaller of this later iteration as we can no longer use the approximation $||x_{(T^o)^c}||_2 = 0$, this affords a larger upper bound on $(***)$ which depends entirely on the input signal. This means a larger numerator for $(\#)$ and similarly we can show the denominator is smaller too with non zero $||x_{(T^o)^c}||_2$. This permits a smaller λ and some means a lesser probability of ϵ_{cv}^o being smaller than ϵ_{cv}^p. We can chain such reasoning to say that the effect of $||x_{(T^o)^c}||_2$ is removed when this term is finally zero and hence CV output with high probability accomplishes the entire support set.
If even if all our iterations do not recover entire support set and oracle output was answer of last iteration, then we can say that the unrecovered columns were basically treated as noise, and the entire argument can be repeated with a different σ_n which also includes the term $||x_{T \setminus T^o}||$. We should increase the number of iterations , or add more measurements.

3 Johnson–Lindenstrauss (JL) lemma

The lemma states that for an accuracy paramter $\epsilon \in (0, 0.5]$ and confidence parameter $\delta \in (0,1)$ and an integer $r \geq r_0 = \lceil C\epsilon^{-2} log(1/2\delta) \rceil$. Let M be a r x N matrix of iid random variables R such that

1. $Var[R] = 1/r$
2. $E(R) = 0$
3. $P(|R| > \lambda) \leq 2e^{-a\lambda^2}$ for fixed a and all λ, i.e a tighter tail bound than Gaussian.

Col 3.1: Then for an $x \in R^N$ with probability $1-\delta$,

$$(1-\epsilon)\|x\|_2 \leq \|Mx\|_2 \leq (1+\epsilon)\|x\|_2$$

Now for a fixed x, a random matrix M fails with probability $\leq \delta$. For p different (but fixed x) it fails for at least one with probability $\leq p\delta$ as $P(\cup A_i) \leq \sum_i P(A_i)$. Let $\theta = p\delta$, then the bounds of JL Lemma holds with probability $\geq 1-\theta$ for all fixed vectors $x_1, x_2, ..., x_p$.

This lemma helps us to decide the number of cross validation measurements.

3.1 Useful Notations and Properties

Let $a \sim_\epsilon b$ denote $(1-\epsilon)a \leq b \leq (1+\epsilon)a$. Some properties are

1. For positive a, b if $a \sim_\epsilon b$, then $b/[(1+\epsilon)(1-\epsilon)] \sim_\epsilon a$
2. If $a \sim_\epsilon b$ and $c \sim_\epsilon d$, then $a(1-\epsilon)/[c(1+\epsilon)] \leq b/d \leq a(1+\epsilon)/[c(1-\epsilon)]$. We can show $(1-\epsilon)/(1+\epsilon) \geq (1-3\epsilon)/(1-\epsilon)$ so $a/c \sim_\delta b/d$ for $\delta = 2\epsilon/(1-\epsilon)$
3. Let amongst $a_1, a_2, .., a_p$, a_l be minimum and among $b_1, b_2, .., b_p$, b_k be minimum with $a_i \sim_\epsilon b_i$ for $i \in [1, p]$. Now $b_k \geq a_k(1-\epsilon) \geq a_l(1-\epsilon)$ and $b_k \leq b_l \leq a_l(1+\epsilon)$, so $min_j(a_j) \sim_\epsilon min_j(b_j)$

3.2 Proxies for Error Terms

For p vectors $u_j = x - \hat{x}_j$ where $\hat{x}_j$ is the recovered signal in the j^{th} iteration, we apply *Col 3.1* to these p vectors to get for $j = 1, 2, ...p$,

$$\frac{1}{1+\epsilon} \leq \frac{\|u_j\|_2}{\|\Psi u_j\|_2} \leq \frac{1}{1-\epsilon} \tag{3.2.1}$$

We can take $r = r_0$ with $\delta = \theta/p$ and choose θ for our usage. C can be taken as 8 for Gaussian matrices since $C \geq ln(4)/(4a)$ and $a = 1/2$ by using Gaussian Tail Bounds and ignoring the $\sqrt{r}$ term. We can use 3.1.3 to get

$$\frac{1}{1+\epsilon} \leq \frac{min_j\|u_j\|_2}{min_j\|\Psi u_j\|_2} \leq \frac{1}{1-\epsilon} \tag{3.2.2}$$

wherein $min_j\|u_j\|_2$ is the oracle error and $min_j\|\Psi u_j\|_2$ is the CV error.

We have $\|x\|_2 \sim_\epsilon \|\Psi x\|_2$ and $\|x - x_j\|_2 \sim_\epsilon \|\Psi(x - x_j)\|_2$. Use 3.1.2 to get

$$1 - \delta \leq \frac{\|x - x_j\|_2/\|x\|_2}{\|\Psi(x - x_j)\|_2/\|\Psi x\|_2} \leq 1 + \delta \tag{3.2.3}$$

where δ is defined in 3.1.2

4 Applications to CS

4.1 Bounding error in k sparse approximation

We know that if $m \times N$, matrix Φ satisfies $2k$-RIP with parameter δ and $\delta < 0.41$, then the error between x and the approximation $\hat{x} = L_1(\Phi, \Phi x) = \text{argmin}_{\Phi x = \Phi z} \|z\|_1$, is given by

$$\|x - \hat{x}\|_2 \leq \frac{c}{\sqrt{k}}\sigma_k(x) = \frac{c}{\sqrt{k}}\|x - x_k\|_1 \tag{4.1.1}$$

where $x_k \in R^n$ is the k sparse vector that corresponds to the vector x with all but the largest k entries set to zero.

We can use cross validation to obtain bounds on the quantity $\sigma_k(x)$. This is useful to tell us how nicely can a k sparse vector approximate x. Using AM-AGM and JL Lemma,

$$(1-\epsilon')\|\Psi(x-\hat{x})\|_2 \leq \frac{c}{\sqrt{k}}\sigma_k(x) \leq c\|x-x_k\|_2$$

Also we know that in 3.3.1 we can replace $\hat{x}$ with the best k sparse approximation of $\hat{x}$ and still have the bounds maintained with c replaced by $3c$. We change $\hat{x}$ to $\hat{x}_k$, the best k sparse approximation to $\hat{x}$ in the above discussion. Thus we also get

$$\|x-\hat{x}_k\|_2 \geq \|x-x_k\|_2$$

since x_k is the best approximation to x Now again use JL Lemma to get

$$(1+\epsilon')\|\Psi(x-\hat{x})\|_2 \geq \|x-\hat{x}_k\|_2 \geq \|x-x_k\|_2$$

Thus as we see we have bounded the error in approximating x with a k-sparse vector using a cross validation matrix Ψ and CS decoding algorithms. If we have a series p of p sparsity levels where we want to bound the in approximating x, then Ψ can have around $10logp$ rows.

4.2 Number of measurements

Natural images have compressible wavelet sequences $x \in R^n$ which obey a power law decay,

$$|x|_{(k)} \leq c_s k^{-s}$$

where $x_{(k)}$ is the kth largest coefficient in x, $s > 1$ is the level of compressibility and c_s is a constant that depends on s and normalization of x. Then we can see for compressible signals

$$\|x-x_k\| = c_s(\sum_{i=1}^{k}(k+i)^{-s} = c_s(\sum_{i=1}^{k}(k+i)^{-s} + \sum_{i=k+1}^{3k}(k+i)^{-s}...)$$

$$\leq c_s(\sum_{i=1}^{k}k^{-s} + \sum_{i=k+1}^{3k}(2k)^{-s}...) = c_s(k*k^{-s} + 2k*(2k)^{-s} + 4k*(4k)^{-s}...)$$

$$= c_s * k^{-s+1}(1+2^{-s+1}+4^{-s+1}...) \leq c_s' * k^{-s+1}$$

where $c_s' = c_s * (1+2^{-s+1}+4^{-s+1}...)$ where the second term is bounded because it represents the sum of a GP with common ratio < 1. So,

$$\|x-x_k\|_1/\sqrt{k} \leq c_s' * k^{-s+\frac{1}{2}}$$

So, for $\hat{x_m} = L(\Phi, \Phi x)$ with Φ as an $m \times N$ matrix satisfying $2k$RIP with parameter δ

$$\|x-x_k\|_2 \leq c_{s,\delta} * k^{-s+\frac{1}{2}}$$

The number of measurements requires to achieve a desired upper bound on $\|x-x_k\|_2$ will therefore depend on the sparsity and compressibility of the image and can be estimated adaptively using cross validation

4.3 Choice of Regularisation parameters

Several reconstruction algorithms involve calculating several intermediate estimates x_j of the possible solution. For the LASSO problem, the objective function is

$$x^{\lambda} = \text{argmin}_{z \in R^N} \|\Phi x - \Phi z\|_2 + \lambda \|z\|_1$$

where the second term enforces the sparsity constraints.
The homotopy continuation algorithms is one of the algorithms used to find the value of x^{λ} at a predetermined value of λ. Thus, we start with a sufficiently large value of λ so that x^{λ} is effectively 0 and we decrease it to reach the required level. Note that for $\lambda = 0$ the case is reduced to the solution of the l1 minimizer: $\text{argmin}_{\Phi x = y} \|z\|_1$. For $\lambda = 0$ it is a vector that makes the fidelity term 0. At $\lambda = 0^+$, the minimization occurs again at one of such vectors however we pick the one with with l_1 norm to get the minimum since λ is small however non zero. Since the solution path(as a function of λ) is continuous, the solution at 0 is the vector with smallest l_1 norm making the fidelity term 0. It can be proved that for this algorithm the solution path (i.e., variation of $x^{\lambda} \in R^n$ with λ) is piecewise-affine function with kinks at finite $\lambda \in \lambda_1, ..., \lambda_p$. Now we can apply cross-validation using these points to get an estimate of the optimal x for the given objective function and get an appropriate value of the optimal λ from the sequence $\lambda_1, ..., \lambda_p$. To this end, we use the equation (3.2.2), which states,

$$\frac{1}{1+\epsilon} \leq \frac{min_{\lambda}\|x - x^{\lambda}\|_2}{min_{\lambda}\|\Psi(x - x^{\lambda})\|_2} \leq \frac{1}{1-\epsilon}$$

Thus using an appropriate CV matrix Ψ we can estimate the minimum error to be within an ϵ bound of $min_{\lambda}\|\Psi(x - x^{\lambda})\|$.
Now, since the kinks occur when the sign of an element of x changes in the solution path or some non zero element becomes zero or vice-versa, there can be at most 3^N kinks. This gives us an $O(N)$(using JL Lemma) bound on the size of Ψ which defeats the purpose of compressed sensing. However, for m measurements we observe that we get $O(m)$ kinks generally. Thus, we can heuristically use $O(logm)$ measurements out of m in the CV matrix to ensure that the recovery error and the CV error are within $epsilon$ distance of each other. Thus, we find $x^{\lambda} = \text{argmin}(\Psi(x - \hat{x}^{\lambda})$ and for this value of λ use all the m measurements to get the final approximation of x.

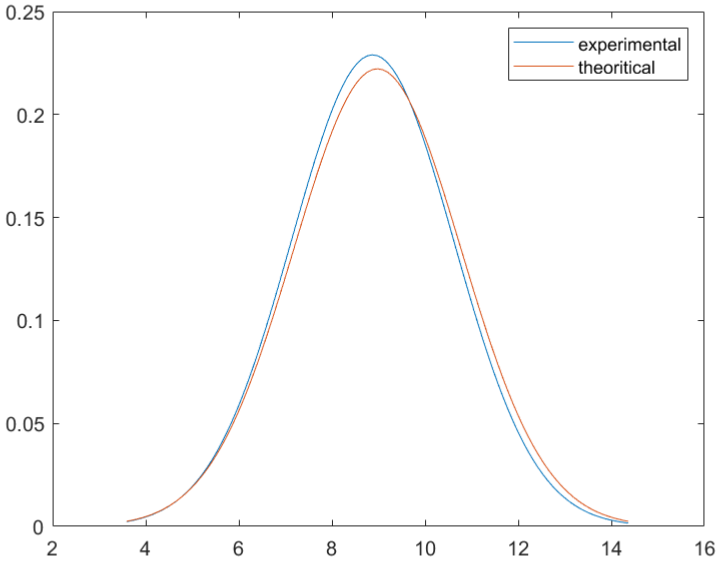
4.4 Choice of sparsity parameter

In greedy algorithms like OMP, a sequence intermediate solutions $\hat{x_j}$ are generated till after k iterations a $k$sparse vector is returned, where k is an input to the algorithm. OMP will recover with high probability a vector x having at most $k \leq m/\log N$ nonzero coordinates from its image Φx if Φ is a (known) $m \times N$ Gaussian or Bernoulli matrix with high probability. For the general vector $x = x_d + N$ where N is the noise vector, an intermediate estimate $\hat{x_s}$ is expected to be a better estimate of x then x_k if $d << k$. Thus, we can use the sequence of intermediate solutions to estimate the noise level using cross validation.

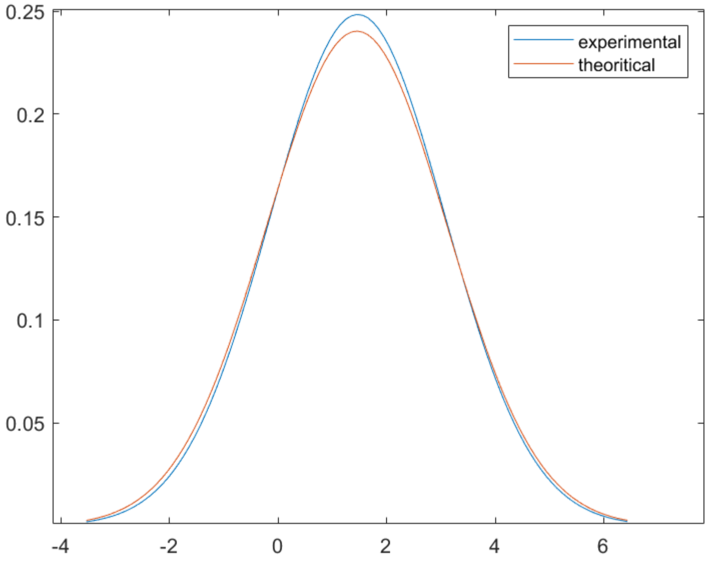
5 Simulations

5.1 Probabilistic simulations

1. We show ϵ_{cv} follows the distribution we proved.



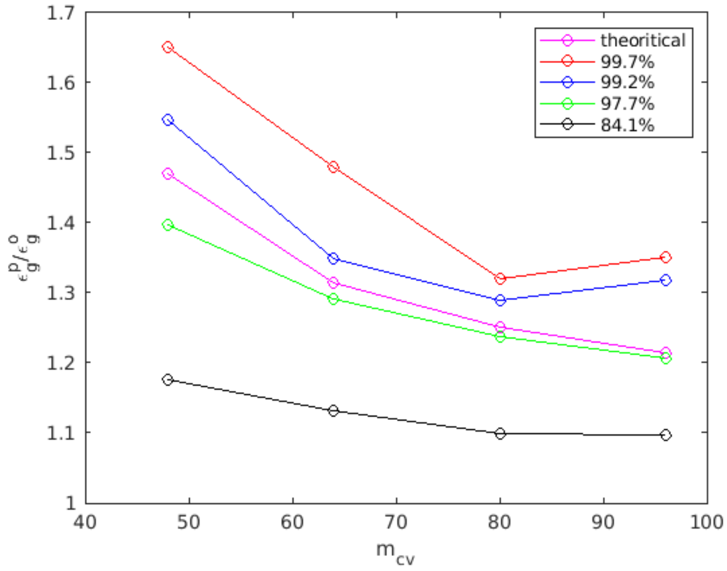
2. We show $\Delta\epsilon_{cv}$ follows the distribution we proved.



5.2 Simualtion of Section 2.4

We run OMP-CV a lot of times and see what how many times were we able to recover the entire support set. According to section 2.4, with high probability generalised oracle error bounds OMP-CV output oracle error to a constant factor. This we depict here by running OMP-CV 2000 times for different m_cv and see the various percentiles of the ratio $\epsilon_g^p/\epsilon_g^o$. The parameters are

1. signal size = 1000
2. number of measurements=400
3. sparsity=50
4. OMP-CV iterations=150



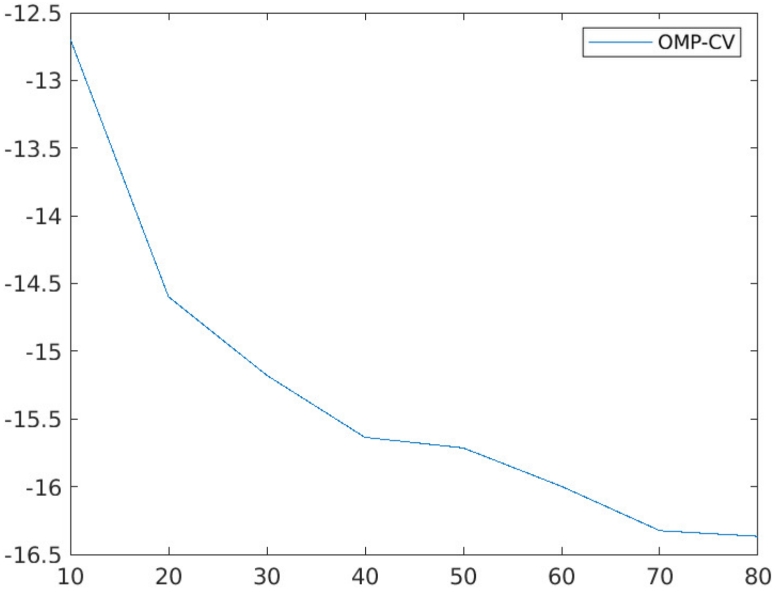
5.3 Simulation for OMP-CV algorithm

1. signal size = 1000
2. sparsity=50
3. OMP-CV iterations=100

5.3.1 On number of CV measurements

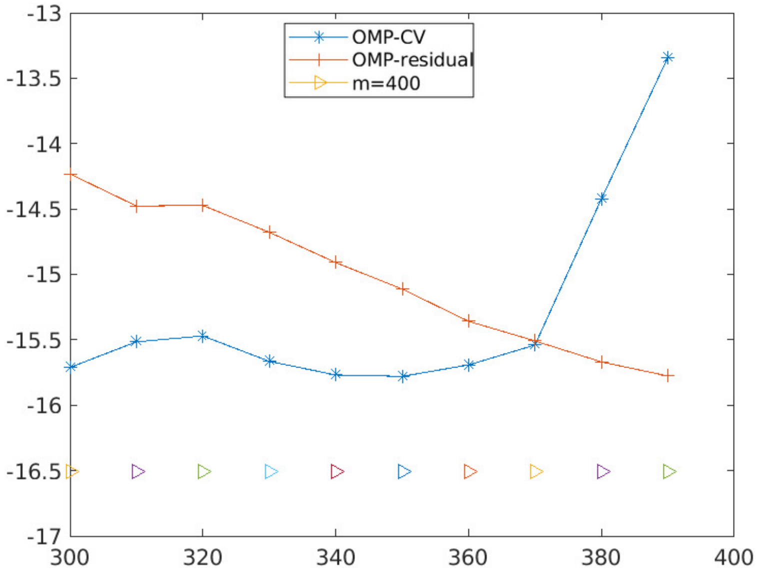
The algorithm is run 300 times for each of the m_{cv} values ranging from 10 to 80 with noise value $\sqrt{0.1}$. The log of the average of the recovery error values are plotted against m_{cv}. The reconstruction gets better with increasing values of m_{cv} indicating that for the same m for the matrix A increasing the CV measurements improves performance.

However the rate decreases with increase in value of m_{cv} showing that after a fixed value the improvement would be negligible. Hence we do not need to have very large values of m_{cv} for decent performance.



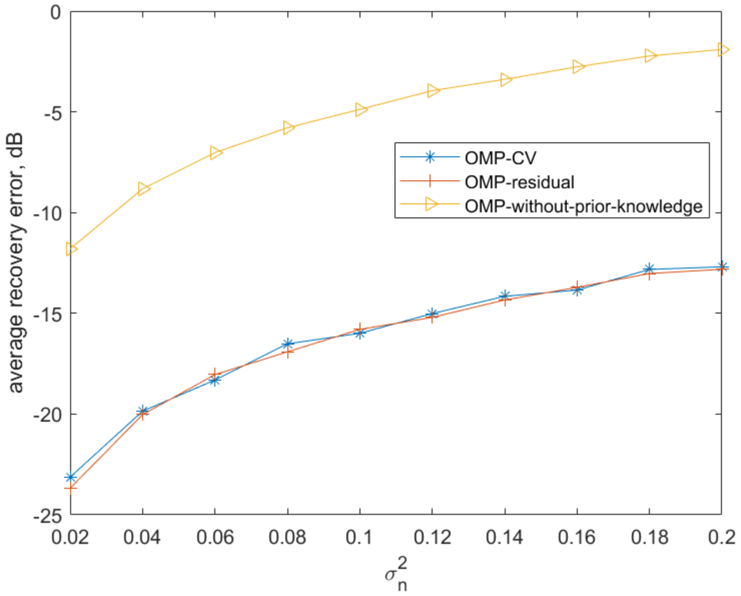
5.3.2 Trade off between m and mcv

We study the variation in recovery error when sum of m and m_{cv} is kept fixed $= 400$ For this measurement, we run both the OMP-CV and OMP-residual (with noise level based termination) 300 times with noise $\sqrt{0.1}$. The plot of the log of average recovery error is plotted alongside the m values. We see that OMP-CV is almost always better than OMP-residual except when m_{cv} is very small. Also plotted is the OMP-residual output for $m = 400$ and it is quite close the OMP-CV output



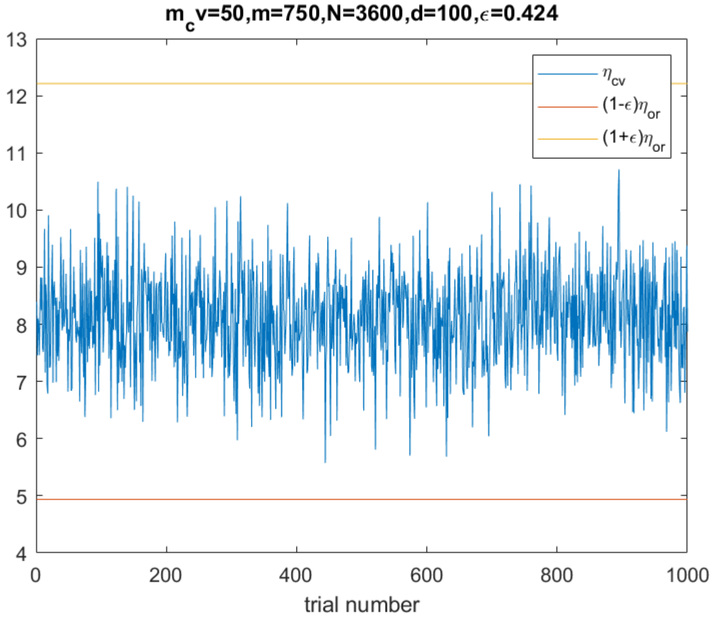
5.3.3 On noise level

We study the effect of increasing noise variance of the data. We run OMP-residual, OMP-CV and OMP without prior knowledge of noise (i.e., with termination condition as $\|y - Ax\| < 10^{-5}\|y\|$. As expected the performance deteriorates as noise level increases. Also, OMP-CV and OMP-residual(which uses the noise variance) show similar performance which is far superior to that of OMP without prior knowledge. Thus, we can say that OMP-CV performs better when noise level is not known.

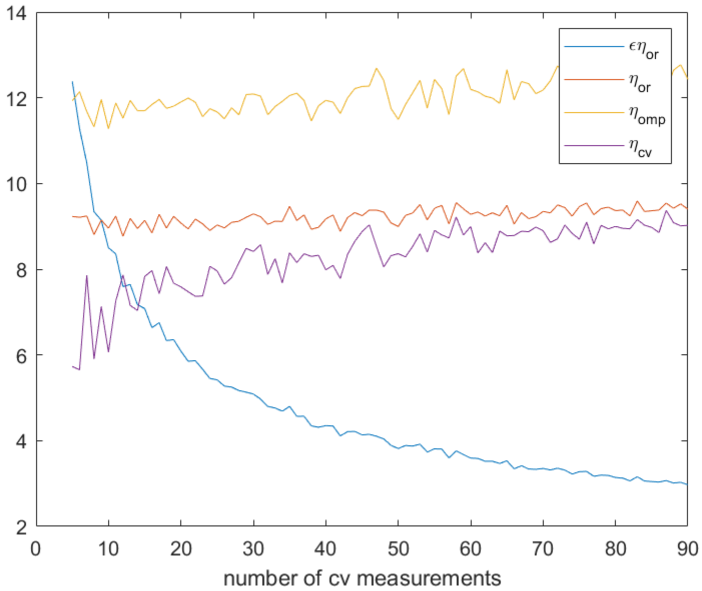


5.4 Simulations of OMP

The script SectionVI_first_sim.m realises random CV matrices and shows that oracle error $\sim_\epsilon$ CV error.



The script SectionVI_second_sim.m shows that as $r = m_{cv}$ increases the bound on difference between CV error and oracle error gets smaller($= \epsilon\eta_{cv}$). CV error becomes closer to oracle error and is hence a better proxy. Also CV error is always below OMP error, hence doing cross validation is always superior.



5.5 CS Decoding with adaptive measurements

In the following algorithm, we successively decrease measurements given to CV and give them to OMP.
The equation (27) serves as a proxy for RMSE of the recovered signal. The algorithm returns when (27) is below a user defined threshold. If that threshold is not reached we return the estimate using all the measurements.

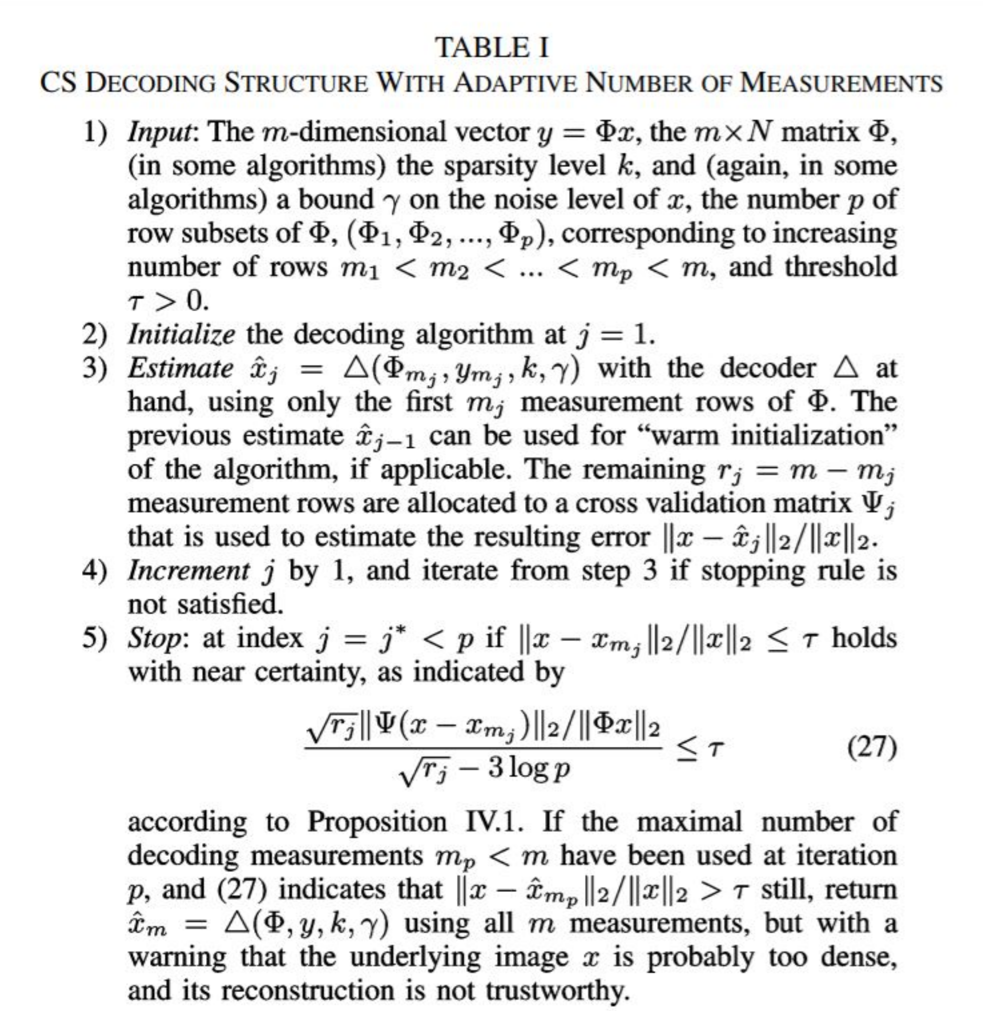
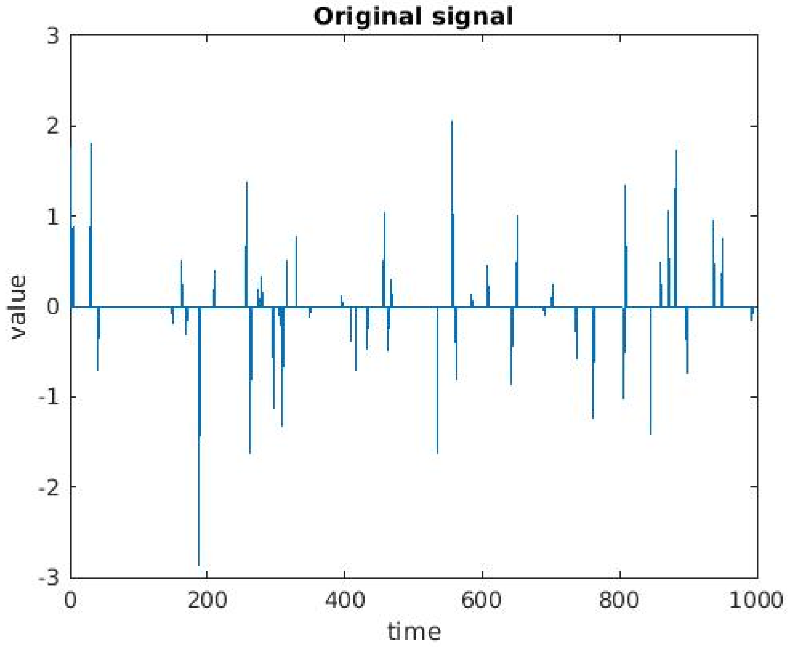


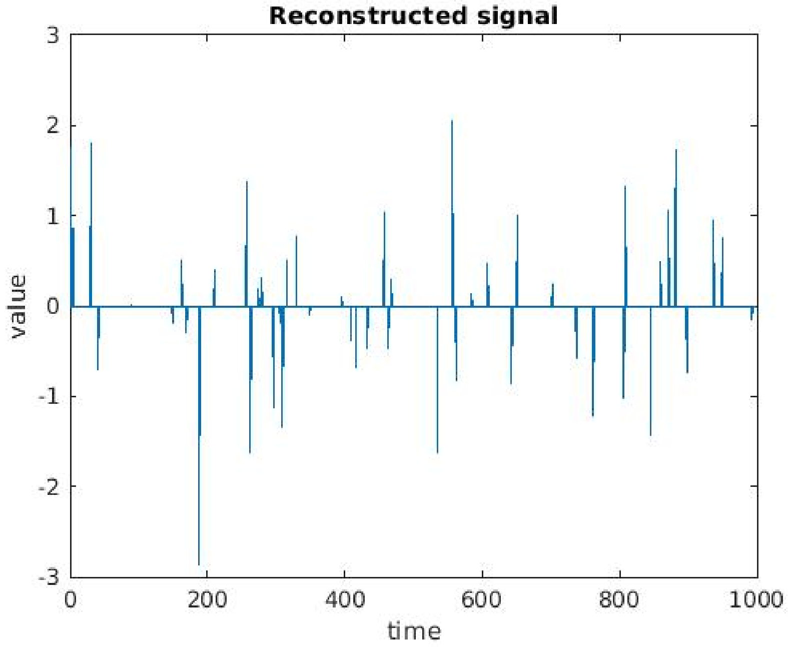
TABLE I
CS DECODING STRUCTURE WITH ADAPTIVE NUMBER OF MEASUREMENTS

1) *Input*: The m-dimensional vector $y = \Phi x$, the $m \times N$ matrix Φ, (in some algorithms) the sparsity level k, and (again, in some algorithms) a bound γ on the noise level of x, the number p of row subsets of Φ, $(\Phi_1, \Phi_2, ..., \Phi_p)$, corresponding to increasing number of rows $m_1 < m_2 < ... < m_p < m$, and threshold $\tau > 0$.
2) *Initialize* the decoding algorithm at $j = 1$.
3) *Estimate* $\hat{x}_j = \triangle(\Phi_{m_j}, y_{m_j}, k, \gamma)$ with the decoder $\triangle$ at hand, using only the first m_j measurement rows of Φ. The previous estimate $\hat{x}_{j-1}$ can be used for "warm initialization" of the algorithm, if applicable. The remaining $r_j = m - m_j$ measurement rows are allocated to a cross validation matrix Ψ_j that is used to estimate the resulting error $\|x - \hat{x}_j\|_2 / \|x\|_2$.
4) *Increment* j by 1, and iterate from step 3 if stopping rule is not satisfied.
5) *Stop*: at index $j = j^* < p$ if $\|x - x_{m_j}\|_2 / \|x\|_2 \leq \tau$ holds with near certainty, as indicated by

$$\frac{\sqrt{r_j} \|\Psi(x - x_{m_j})\|_2 / \|\Phi x\|_2}{\sqrt{r_j} - 3 \log p} \leq \tau \qquad (27)$$

according to Proposition IV.1. If the maximal number of decoding measurements $m_p < m$ have been used at iteration p, and (27) indicates that $\|x - \hat{x}_{m_p}\|_2 / \|x\|_2 > \tau$ still, return $\hat{x}_m = \triangle(\Phi, y, k, \gamma)$ using all m measurements, but with a warning that the underlying image x is probably too dense, and its reconstruction is not trustworthy.

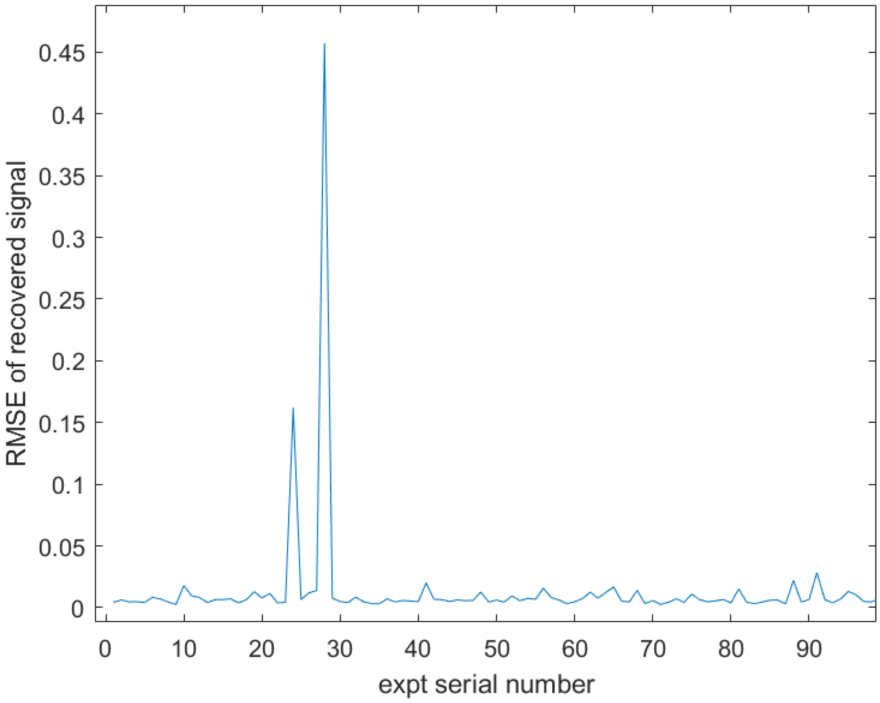




5.6 Homotopy continuation with cross validation

We use the algorithm in the paper *homotopy continuation for sparse signal representation* by *Dmitry M. Malioutov, Mujdat Cetin, and Alan S. Willsky* to solve the LASSO problem. The implementation of the algorithm was taken from here.
We get different recovered signals corresponding to various lambda and then apply CV

to choose a lambda. For this lambda we use all the measurements to get a final estimate of the signal.


RMSE of recovered signal
expt serial number