Assignment 4

Sagar Poudel - 213051001
Sejal Upadhye - 213050017

March 2022

1

Given a compressed sensed in the form $y = \Phi x + \eta$ where $\mathbf{y}$ the measurement vector, has m elements and Φ is the $m \times n$ sensing matrix.

One way to solve the problem is to solve the LASSO problem. To solve the problem, we will use the Cross validation approach to find the value of λ

The validation error is given as:

$$V(E)(g) = \sum_{i \in V} (y_i - \Phi^i x_g)^2 / |V|$$

where, x_g is the recovered x using respective value of lambda.

Following values are used for the reconstruction:

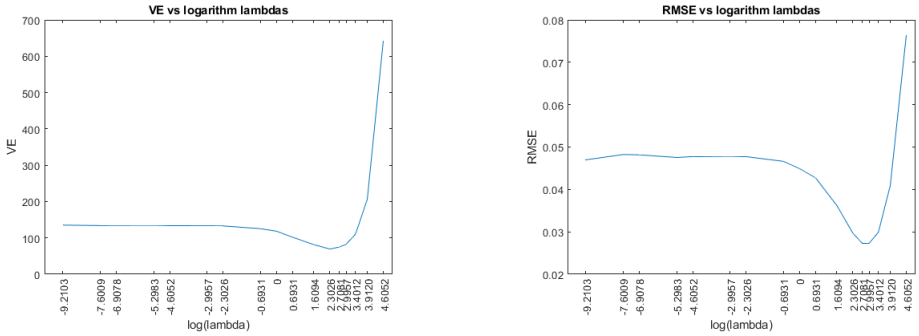
$$n = 500, m = 200, ||x||_0 = 18, \sigma = .05 \times \sum_{i=1}^{m} |\Phi^i x|/m$$

At the first step we used the following set of lambda values:

$$0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10, 15, 20, 30, 50, 100.$$

1.a

On running with subset of the Lambda value above: $0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10$



VE vs $\log(\Lambda)$ RMSE vs $\log(\Lambda)$

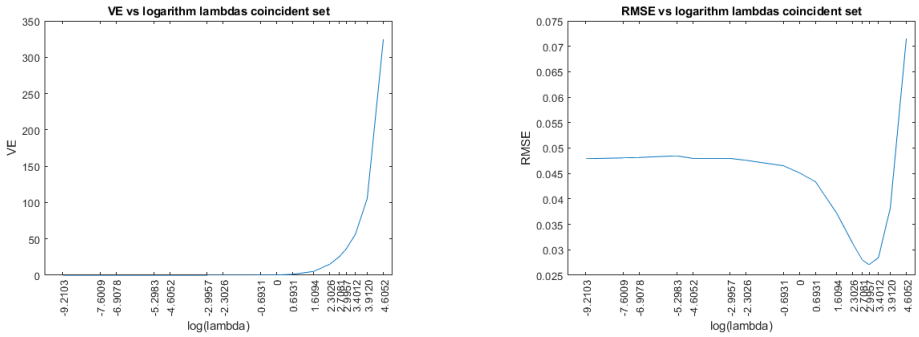
Figure 1: Plot w.r.t. to different value of λ

From the graph we can see that initially the value of VE reduces on increasing the value of lambda, but then after a certain point, the value of VE starts increasing with the increase in value of lambda.

We can also see a similar fashion in RMSE plot with respect to lambda plot. Thus RMSE and VE agree approximately.

1.b

On running with subset of the Lambda value above: $0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2, 5, 10$



VE vs log(Λ) coincidental RMSE vs log(Λ) coincidental

Figure 2: Plot w.r.t. to different value of λ with coincidental set

We can observe similar result but the error is less compare to part a.

1.c

Lemma 1. provides the same proxying ability. Given, Let $\hat{x}$ be a recovered signal and ϵ_x be its recovered error, Provided that m_{cv} is sufficient large, then

$$\in_{cv} = ||y_{cv} - A_{cv}\hat{x}||_2^2 \sim \mathbb{N}(\mu, \sigma^2),$$

where $\mu = \frac{m_{cv}}{m}(\epsilon_x + \sigma_n^2)$, and $\sigma^2 = \frac{m_{cv}}{m^2}(\epsilon_x + \sigma_n^2)^2$

From the above error we can see that, instead of comparing with original signal, it compares the reconstructed signal with y by multiplying with measurement matrix. Likewise with provided that m_{cv} sufficiently large, it can be used in one approximation step of the proof which uses the Central Limit Theorem. Thus it provides the proxying ability.

1.d

(b) Given a regularization parameter $\lambda_N \geq 2\|\mathbf{X}^T\mathbf{w}\|_\infty/N > 0$, any estimate $\widehat{\beta}$ from the regularized lasso (11.3) satisfies the bound

$$\|\widehat{\beta} - \beta^*\|_2 \leq \frac{3}{\gamma}\sqrt{\frac{k}{N}}\,\sqrt{N}\,\lambda_N. \qquad (11.14b)$$

In the above choice, there is a chance that the error bound will go above the bound, if the lambda condition doesn't satisfy. It also gives the error bound but may not give the possible minimum error value for particular data. In CV is a data driven approach, where we can consider λ as hyper-parameter, where we can tune it and choose the one which give less error.

2

2.a

We know that Dictionary representation can express a large number of images into small atom i.e given

$$I = Dx$$

where $I \in classS$ where, D is a dictionary, f is signal and x is coefficient vector and,

$$D \in \boldsymbol{R}^{n\times k}, x \in \boldsymbol{R}^{k\times 1}, I \in \boldsymbol{R}^{n\times 1}$$

Class S1 which consists of images obtained by applying a known derivative filter to the images in S.
Lets filter be denoted by **F'**

$$I' = F'I$$

where F' is filter matrix of $n \times n$ where each row of the dimension generates the new filtered value for each pixel of I

Thus we can write,

$$D' = F'D$$

2.b

Let S_α, S_β be the subset of S.
Applying rotation to new subset we get,

$$I_{2\alpha} = R_\alpha I, \& I_{2\beta} = R_\beta$$

$$I_{2\alpha} = R_\alpha D x_\alpha, \& I_{2\beta} = R_\alpha D x_\beta$$

Using the same linear transformation as question a, we get 2 new dictionary representation
Thus,

$$D_2 = [D_\alpha | D_\beta] \text{ where } D_\alpha = R_\alpha D, \& D_\beta = R_\beta D$$

2.c

Let intensity transformation be denoted as $I_{new} = \alpha I_o ld^2 + \beta I_o ld + \gamma$
I can be expressed as $I = Dx = \sum_{j=1}^{s} d_{mj} x_j$
Now applying I in above transformation we get,

$$I_{new} = \alpha\big(\sum_{j=1}^{s} d_{mj}^2 x_j^2 + 2\sum_{i\neq j}^{s} d_{mj} d_{mi} x_j x_i\big) + \beta(\sum_{j=1}^{s} d_{mj} x_j) + \gamma;$$

Now, observing the above equation, we can tell that new D_3 will be the concatenation of following 4 dictionaries

$$D = [D_s q | D_{p\alpha} | D_\beta | D_\gamma]$$

where,

$$D_s q = \{\alpha d_1^2, \alpha d_2^2, \alpha d_3^2 ..... \alpha d_k^2\}$$
$$D_\alpha = \{....., \alpha d_i . \alpha d_j .....\}$$
$$D_\beta = \{\beta d_1, \beta d_2, \beta d_3 ..... \beta d_k\}$$
$$D_\gamma = ones(\#ofcolumns(D), 1)$$

2.d

This part is same as part a, where blur kernel is defined as liner transformation of an image.

Thus

$$I' = F'I = F'Dx$$

where F' is a blur kernel thus new $D_4 = F'D$

2.e

Let the blur kernel which is linear combination of k set of kernel be F'. It can be defined as:

$$F' = \sum_{i=1}^{k} c_i b_i$$

Now we get the new image by

$$I_{new} = \sum_{i=1}^{k} c_i b_i(I)$$

Thus, using the observation we did for previous part, the final dictionary will be the concatenation of the k blur kernal dictionary

$$i.eD_5 = [b_1(D)|b_2(D).....|b_k(D)]$$

2.f

Let R be the radon transformation of an image in a known angle θ of image I.

The radon transform produces the m bins

Here,

$$I_{new} = RI_{old}, \text{ where } R = n\boldsymbol{R}^{m\times n}, I_{old} \in \boldsymbol{R}^{n\times 1}$$

$$I_{new} = RDx_{old} = (RD)x_{old} = D_6 x_o ld$$

Thus,

$$D_6 = RD, \text{ where R is the radon transform of an image with fixed angle } \theta$$

2.g

Given an image $I \in \boldsymbol{n} \times \boldsymbol{n}$, let a_k be the subset of an image I where $a_k \in \boldsymbol{R}^{n\times k}$ and b_k be the subset of an image I where $b_k \in \boldsymbol{R}^{n\times k}$

$$I_{new} = F' I_o ld = F' Dx$$

where F' is shifting parameter,

When linear shifting is done on above two sets, with zero padding for new value we get different dictionary representation for each subsets. For a_x subset let dictionary denotes as $D_a x$, and b_x subset let dictionary denotes as $D_b x$

$$I_{new} = D_a xax + D_b xb_x$$

Thus $D_6 = [D_a x|D_b x]_{n\times 2k}$

3

3.1

Given:

$$J(\boldsymbol{A_r}) = ||\boldsymbol{A} - \boldsymbol{A_r}||_F^2$$

The above optimization problem is nothing but a solution for low rank matrix approximation of matrix **A**.The result is referred to as the matrix approximation lemma or **Eckart–Young–Mirsky theorem.**

$$\boldsymbol{A} = \boldsymbol{USV}^T \text{where } \mathbf{A} \in \boldsymbol{R}^{m\times n}, n \leq m$$

This can be solved using top r-largest singular values and reconstruct A_r

$$\boldsymbol{A_r} = \boldsymbol{U_1 S_1 V_1}^T$$

where the

$$\boldsymbol{U_1} \in \mathbb{R}^{m \times r}, \boldsymbol{S_1} \in \mathbb{R}^{r \times r}, \boldsymbol{V} \in \mathbb{R}^{r \times n}$$

This approximation used in Image Processing for learning the Bases. In particular K- SVD for finding rank 1 matrix that approximate the E_k as close as possible in square Frobenius norm.

3.2

Given

$$J(\boldsymbol{R}) = ||\boldsymbol{A} - \boldsymbol{RB}||_F^2$$

where **R** is an orthogonal matrix, The solution for the given problem is given as:

$$\min ||\boldsymbol{A} - \boldsymbol{RB}||_F^2 = \min \left|\left|(\boldsymbol{A} - \boldsymbol{RB})^\intercal (\boldsymbol{A} - \boldsymbol{RB})\right|\right|_F$$

$$\begin{aligned}
&= \min \left|\left|(\boldsymbol{A}^\intercal - \boldsymbol{B}^\intercal \boldsymbol{R}^\intercal)(\boldsymbol{A} - \boldsymbol{RB})\right|\right|_F \\
&= \min \left|\left|\boldsymbol{A}^\intercal \boldsymbol{A} - 2\boldsymbol{A}^\intercal \boldsymbol{RB} + \boldsymbol{B}^\intercal \boldsymbol{B}\right|\right|_F \\
&= \max \left|\left|\boldsymbol{A}^\intercal \boldsymbol{RB}\right|\right|_F \\
&= \max \left|\left|\boldsymbol{RBA}^\intercal\right|\right|_F \qquad \text{since trace(AB) = trace(BA)} \\
\text{using SVD on } \boldsymbol{BA}^\intercal \quad \boldsymbol{BA}^\intercal &= \boldsymbol{UDV}^\intercal \\
&= \max \left|\left|\boldsymbol{RUDV}^\intercal\right|\right|_F \\
&= \max \left|\left|\boldsymbol{V}^\intercal \boldsymbol{RUD}\right|\right|_F \\
&= \max \left|\left|\boldsymbol{Z(R)D}\right|\right|_F \qquad Z(R) = V^T RU \\
&= \max \sum_i z_{ii} d_{ii} \leq \sum_i d_{ii} \quad (\because \boldsymbol{Z(R)}^\intercal \boldsymbol{Z(R)} = \boldsymbol{I})
\end{aligned}$$

This quantity $Z(R)$ is an orthogonal matrix and thus the expression is maximised when *Z(R)* equals the identity matrix *I*. Thus

$$\mathbf{R} = \mathbf{VU}^T$$

This problem is also known as Orthogonal Procrustes problem.

The solution from the above problem, can be used to learning the dictionary in Image Processing. It can be use in learning the bases with union of orthonormal bases.

4

Title - Group learning using contrast NMF : Application to functional and structural MRI of schizophrenia
Authors - Vamsi K. Potluru and Vince D. Calhoun
Venue and Year of Publication - IEEE International Symposium on Circuits and Systems, 2008
Link - `https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=4541673`

NMF is useful for decomposing high dimensional data sets into a lower dimensional space. The paper used NMF to learn the features of both structural and functional magnetic resonance imaging (sMRI and fMRI) data. NMF can also be applied to perform group analysis of imaging data and we see NMF being applied to learn the spatial patterns which linearly covary among subjects for both sMRI and fMRI. An additional contrast term, called co-NMF, is added to NMF to identify features distinctive between two groups. The results from co-NMF make sense in light of expectations and are improved compared to the NMF results. The method proposed in the paper is general and may prove to be a useful tool for identifying differences between multiple groups.

Non-negative matrix factorization (NMF) is a tool to split the given data matrix into a product of two non-negative matrix factors. This process can be used to identify useful features in the dataset. The potential of NMF to model spatio-temporal features across subjects is explored in this paper. The use of NMF is attractive for analysis as it naturally incorporates the non-negativeness of the images and is quite easily adapted to learn

across subjects. Also, the constraint of independence among the learnt features is not required for NMF. Adding different constraints to NMF is relatively straightforward as can be seen by the wide range of NMF extensions. The main motivation was to find distinctive features by group analysis of MRI data using NMF. The goal was to identify features distinctive between healthy controls and patients who have been diagnosed with schizophrenia using a version of NMF where the authors explicitly optimized for this. This approach was called contrast NMF or co-NMF.

- NMF model

  Given a non-negative matrix X of size $M \times N$, the task is to split it into a product of two non-negative matrices $\mathrm{W} \in \mathrm{R}^{\geq 0, M \times R}$ and $\mathrm{H} \in \mathrm{R}^{\geq 0, R \times N}$. That is,

$$\mathbf{X} \approx \mathbf{WH} \tag{1}$$

  Here, the Euclidean distance is used as the cost function. The function and its update are given by:

$$\begin{aligned} E &= \|\mathbf{X} - \mathbf{WH}\|_F & (2)\\ \mathbf{W} &= \mathbf{W} \otimes \frac{\mathbf{XH}^\top}{\mathbf{WHH}^\top} & (3)\\ \mathbf{H} &= \mathbf{H} \otimes \frac{\mathbf{W}^\top\mathbf{X}}{\mathbf{W}^\top\mathbf{WH}} & (4) \end{aligned}$$

  Here, $\otimes$ is used to represent element-wise multiplication. The division in the above equations is element-wise. The cost function to be minimized is convex in either W or H but not in both. The rank parameter R on which the sizes of both the matrices W, H depend is usually based on prior knowledge of the data being decomposed. In practice, the matrices W, H are initialized to positive random matrices.

- co-NMF model

  This is a variation of standard NMF. The motivation comes from a brain imaging problem where images of two groups namely healthy controls and patients with schizophrenia are given. NMF can be used to represent the data. However, those features are preferred that not only represent the data but also maximize the difference between the two groups thereby identifying the differences. This is accomplished by introducing an additional constraint that the difference of mean activations of two groups in the learnt factor is maximized.

  This is given by the following objective :

$$\begin{aligned} \min_{\mathbf{W},\mathbf{H}_1,\mathbf{H}_2} \frac{1}{2}\|\mathbf{X}_1 - \mathbf{WH}_1\|_F^2 + \frac{1}{2}\|\mathbf{X}_2 - \mathbf{WH}_2\|_F^2 \\ -\|\lambda \otimes (\mu(\mathbf{H}_1) - \mu(\mathbf{H}_2))\|_1 \end{aligned} \tag{5}$$

  The matrices $X1$, $X2$ correspond to the observation data from two groups, W is the common feature space of the groups and $H1$,$H2$ the corresponding activations. The function μ computes the mean activation by taking a matrix of column vectors and producing a single column vector by averaging across the entire group. The following update equations to minimize the objective is used:

$$\begin{aligned} \mathbf{W} &= \mathbf{W} - \eta[-\mathbf{X}_1\mathbf{H}_1^\top - \mathbf{X}_2\mathbf{H}_2^\top \\ &\quad +\mathbf{WH}_1\mathbf{H}_1^\top + \mathbf{WH}_2\mathbf{H}_2^\top] & (6)\\ \mathbf{H}_1 &= \mathbf{H}_1 \otimes \frac{\mathbf{W}^\top\mathbf{X}_1 + ((\lambda \otimes \mathbf{d})\mathbf{1})^-}{\mathbf{W}^\top\mathbf{WH}_1 + ((\lambda \otimes \mathbf{d})\mathbf{1})^+} & (7)\\ \mathbf{H}_2 &= \mathbf{H}_2 \otimes \frac{\mathbf{W}^\top\mathbf{X}_2 + ((\lambda \otimes \mathbf{d})\mathbf{1})^+}{\mathbf{W}^\top\mathbf{WH}_2 + ((\lambda \otimes \mathbf{d})\mathbf{1})^-} & (8) \end{aligned}$$

where 1 is row vector of ones of appropriate dimension, λ is the weight vector and d is the mean(along columns) difference of matrices $H1$, $H2$. We use $\pm$ in the superscript to denote the absolute values of positive and negative elements of matrix with the rest set to zero respectively. The objective function is also not scale free. Therefore, gradient descent is used for updating W and then rescale its column vectors to norm unity. Multiplicative update rules for W could have been employed.

5

Given,

$$\boldsymbol{y} \sim \text{Poisson}(I_o \exp(-\boldsymbol{R}\boldsymbol{f}))$$

where $\boldsymbol{y} \in \mathbb{R}^m$,
m = number of projection angles $\times$ number of bins per angle;
I_o is the power of the incident X-Ray beam;
$\boldsymbol{R}$ represents the Radon operator (effectively a $m \times n$ matrix)
$f \in \mathbb{R}^n$ is signal(tissue density values)

Let consider the single row for the calculation

$$\boldsymbol{y_i} \sim \text{Poisson}(I_o \exp(-\boldsymbol{R^i}\boldsymbol{f}))$$

For the good estimate of **f** we will take log likelihood of the above distribution

$$-log\big(\ P(y_i|R^i f)\big) = -log\big(\frac{(I_o e^{-R^i f})^{y_i} e^{-I_o e^{-R^i f}}}{(y_i)!}\big)$$

f is sparse in standard DCT basis, thus our change become:

$$-log\big(\ P(y_i|R^i \Psi\theta)\big) = -log\Big(\frac{(I_o e^{-R^i \Psi\theta})^{y_i} e^{-I_o e^{-R^i \Psi\theta}}}{(y_i)!}\Big)$$

$$= -y_i log(I_o e^{-R^i \Psi\theta}) + I_o e^{-R^i \Psi\theta} + log((y_i)!)$$

$$= -y_i log(I_o) + y_i R^i \Psi\theta + I_o e^{-R^i \Psi\theta} + log((y_i)!)$$

Now taking for all m,

$$\sum_{i=1}^{m} -y_i log(I_o) + y_i R^i \Psi\theta + I_o e^{-R^i \Psi\theta} + log((y_i)!)$$

Now finally Optimization become,

$$\sum_{i=1}^{m} -y_i log(I_o) + y_i R^i \Psi\theta + I_o e^{-R^i \Psi\theta} + log((y_i)!) + \lambda||\theta||_1$$

where last term is for regularization sparsity constraint. Only considering θ then, removing the other value we get,

$$J(\theta) = \sum_{i=1}^{m} y_i R^i \Psi\theta + I_o e^{-R^i \Psi\theta} + \lambda||\theta||_1$$

Suppose, there is another noise i.e Gaussian noise presents, We get,

$$y = \text{Poisson}(I_o \exp(-\boldsymbol{R}\boldsymbol{f})) + \text{Gaussian}(0, \sigma)$$

$$y = p + \eta, \text{ where } \eta \sim N(0, \sigma)$$

Now considering the only one row we get, taking the -ve log likelihood. we get,

$$-p_i log(I_o e^{-R^i \Psi\theta}) + I_o e^{-R^i \Psi\theta} + log((p_i)!) + \frac{y_i - v_i}{2\sigma^2}$$

Considering all m, we get

$$\sum_{i=1}^{m} -p_i log(I_o e^{-R^i \Psi \theta}) + I_o e^{-R^i \Psi \theta} + log((p_i)!) + \frac{y_i - v_i}{2\sigma^2}$$

Adding new regularization sparsity constraint we get,

$$J(\theta, v) = \sum_{i=1}^{m} \Big(- p_i log(I_o e^{-R^i \Psi \theta}) + I_o e^{-R^i \Psi \theta} + log((p_i)!) + \frac{y_i - v_i}{2\sigma^2} \Big) + \lambda ||\theta||_1$$

The above problem can be solved by solving the value of θ and v by taking the derivative with respect to respective entity.