Assignment 3: CS 754

Md Kamran, Samiksha Das

Due 23 March 2021

Contents

Question 1

a)The restricted eigen value condition dictates that

$$\frac{1}{N}\frac{\nu^T X^T X \nu}{\|\nu\|_2^2} \geq \gamma$$

for all nonzero $\nu \in C$ where ν is the pertubation vector, γ is a parameter X is model matrix . This is a specific case of restricted strong convexity in linear regression.

b)Since $\hat{\nu} = \hat{\beta} - \beta*$ minimises $G(\nu)$ by construction,since $\hat{\beta}$ minimises $\frac{1}{2N}\|y - X\beta\|_2^2 + \lambda_N \|\beta\|_1$, $(\hat{\nu})$ is the argmin of $G(\nu) = \frac{1}{2N}\|y - X(\beta * + \nu)\|_2^2 + \lambda_N\|\beta * + \nu\|_1$ hence $G(\hat{\nu}) \leq G(0)$.

c)From above since $G(\hat{\nu}) \leq G(0)$,

$$\implies \frac{1}{2N}\|y - X(\beta * + \hat{\nu})\|_2^2 + \lambda_N\|\beta * + \hat{\nu}\|_1 \leq \frac{1}{2N}\|y - X\beta * \|_2^2 + \lambda_N\|\beta * \|_1$$

Now $y = X\beta * + w$;

$$\implies \frac{1}{2N}\|w - X\hat{\nu}\|_2^2 + \lambda_N\|\beta * + \hat{\nu}\|_1 \leq \frac{1}{2N}\|w\|_2^2 + \lambda_N\|\beta * \|_1$$

$$\frac{1}{2N}\|w - X\hat{\nu}\|_2^2 = \frac{1}{2N}(w - X\hat{\nu})^T(w - X\hat{\nu})$$

$$= \frac{1}{2N}(w^T w - \hat{\nu}^T X^T w - w^T \hat{\nu} X + (X\hat{\nu})^T(X\hat{\nu})) = \frac{1}{2N}(\|w\|_2^2 + \|X\hat{\nu}\|_2^2) + \frac{1}{N}w^T\hat{\nu}X$$

since $w^T\hat{\nu}X$ is a scalar, $w^T\hat{\nu}X = \hat{\nu}^T X^T w$.Then,

$$\implies \frac{1}{2N}(\|w\|_2^2 + \|X\hat{\nu}\|_2^2) + \frac{1}{N}w^T\hat{\nu}X + \lambda_N\|\beta * + \hat{\nu}\|_1 \leq \frac{1}{2N}\|w\|_2^2 + \lambda_N\|\beta * \|_1$$

$$\frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \frac{1}{N}w^T\hat{\nu}X + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

QED

d)Since $1/\infty + 1/1 = 1$, applying Holder's inequality,$\{p = \infty, q = 1\}$ and since $w^T X\hat{\nu}$ is scalar,

$$w^T X\hat{\nu} = \|w^T X\hat{\nu}\| \leq \|w^T X\|_\infty \|\hat{\nu}\|_1$$

Then,applying the above into 11.21,

$$\frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \frac{1}{N}w^T\hat{\nu}X + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\} \leq \frac{1}{N}\|w^T X\|_\infty\|\hat{\nu}\|_1 + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

$$\frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \frac{1}{N}\|Xw^T\|_\infty\|\hat{\nu}\|_1 + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

QED

e)from 11.22,

$$\frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \frac{1}{N}\|Xw^T\|_\infty\|\hat{\nu}\|_1 + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

By assumption $\frac{1}{N}\|X^T w\|_\infty \leq \lambda_N/2$,

$$\implies \frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \lambda_N/2\|\hat{\nu}\|_1 + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

Now $\|\hat{\nu}\|_1 = \|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^C}\|_1$ hence;

$$\implies \frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq \lambda_N/2(\|\hat{\nu}_S\|_1 + \|\hat{\nu}_{S^C}\|_1) + \lambda_N\{\|\beta * \|_1 - \|\beta * + \hat{\nu}\|_1\}$$

Now,

$$\frac{1}{2N}\|X\hat{\nu}\|_2^2 \leq 3/2\lambda_N\|\hat{\nu}_S\|_1 - 1/2\lambda_N\|\hat{\nu}_{S^C}\|_1 \leq 3/2\lambda_N\|_1\hat{\nu}_S\| \leq 3/2\lambda_N\sqrt{k}\|\hat{\nu}\|_2$$

since by RMS-AM inequality, $\|_1\hat{\nu}_S\| \leq \sqrt{k}\|\hat{\nu}_S\|_2 \leq \sqrt{k}\|\hat{\nu}\|_2$.
QED **f)** Lemma 11.1 allows us to apply the γ-RE condition to $\hat{\nu}$

$$\frac{1}{N}(\hat{\nu}^T X^T X\nu)/\|\hat{\nu}\|_2^2 \geq \gamma \implies \frac{1}{N}\|X\nu\|_2^2/\|\hat{\nu}\|_2^2 \geq \gamma$$

Combining this lower bound with earlier inequality (11.23) yields

$$\gamma/2\|\hat{\nu}\|_2^2 \leq 3/2\lambda_N\sqrt{k}\|\hat{\nu}\|_2$$

and rearranging yields the bound (11.14b).
QED

g)$\lambda_N \geq 2\frac{\|X^T w\|_\infty}{N}$ shows up while deriving 11.23 which provides bound for $\frac{1}{2N}\|X\hat{\nu}\|_2^2$ and also while proving lemma 11.1 to arrive at $0 \leq \lambda_N/2\|\hat{\nu}\|_1 + \lambda_N(\|\hat{\nu}_S\|_1 - \|\hat{\nu}_{S^C}\|_1)$.
h) A convex loss function in high-dimensional settings (with p N) cannot be strongly convex; rather, it will be curved in some directions but flat in others. As shown in Lemma 11.1, the lasso error $\cap\nu = \cap\beta\beta*$ must lie in a restricted subset C of R^p. For this reason, it is only necessary that the loss function be curved in certain directions of space.

i)Advantages of given theorem over theorem 3:-
l_2-error decays more quickly for k/N compared to that of theorem-3 and hence the bound by the above theorem is better.
The Restrictive Eigenvalue condition is less restrictive and more general compared to RIP. Hence it applies on larger cases compared to RIP. So the given theorem has a larger range of satisfying matrices.
Since there are more parameters affecting the bound in case of the given theorem, we have a larger freedom to fine tune our parameter λ_N and hence the $l-2$ error can be more optimised.
The bound of the $l-2$ error given by the above theorem is independent of any direct dependence on $l-1$ error as opposed to Theorem-3. This makes it easier to bind the $l-2$ error as the bound only depends tunable and regularisation parameters.

Advantages of theorem 3 over given theorem:-
There are more variables with bounds in the given theorem compared to theorem 3 which bounds on only 2 variables.
In the given theorem we have assumed independence of noise signal.But this might not be the case always. There is no such assumption in theorem 3 which makes it more general.
Theorem 3 depends only on the maximum value of the noise signal where as the given theorem depends on the standard deviation of noise signal.Slight variations in the noise signal are more likely to affect the standard deviation than the maximum value. Hence, Theorem 3 is better for varying noise signal.

j)The similarity between the bounds on 'Dantzig selector' and the LASSO are:-
The parameter λ which is present in the error bounds of both the functions are lower bounded by terms proportional to the $l-\infty$ norm of $X^T w$ and $A^T e$ respectively where X, A are the measurement matrices and w, e are the noise vectors.
Also, the probability of the tail bound of the $l-\infty$ norm of $X^T w$ and $A^T e$ are exponentially decreasing with increasing value of the bound.

k)The square-root LASSO method is pivotal in that it neither relies on the knowledge of the standard deviation nor does it need to pre-estimate . Moreover, the method does not rely on normality or sub-Gaussianity of noise. It achieves near-oracle performance, attaining the convergence rate (s/n) log p1/2 in the prediction norm, and thus matching the performance of the lasso with known . These performance results are valid for both Gaussian and non-Gaussian errors.

Question 2

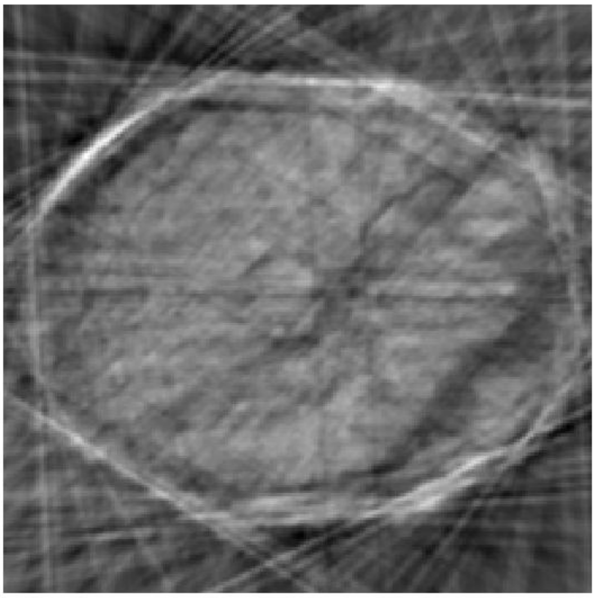

Figure 1: Filtered Back Projection using Ram-Lak Filter

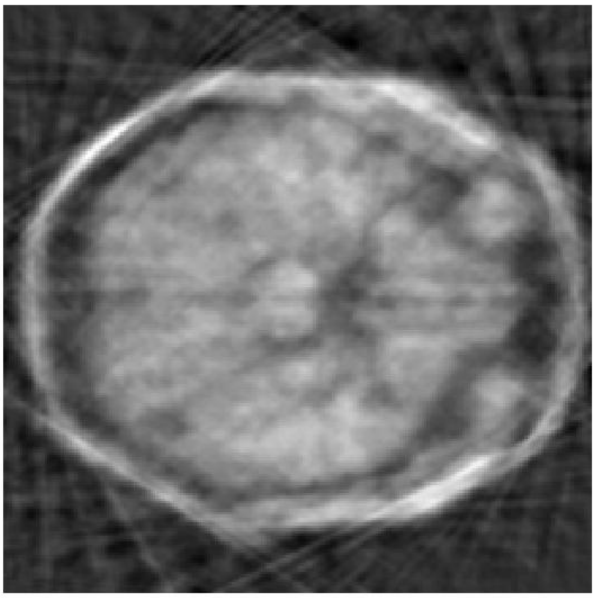

Figure 2: Independent CS-based Reconstruction

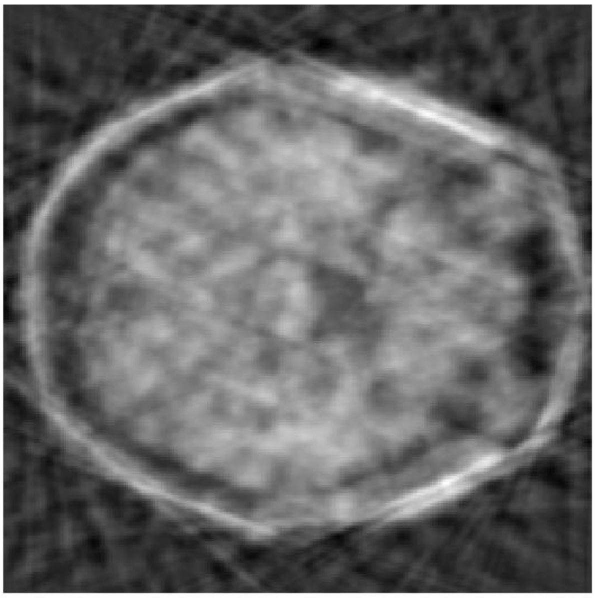

Figure 3: Coupled CS-based Reconstruction using 2 slices, Image 1

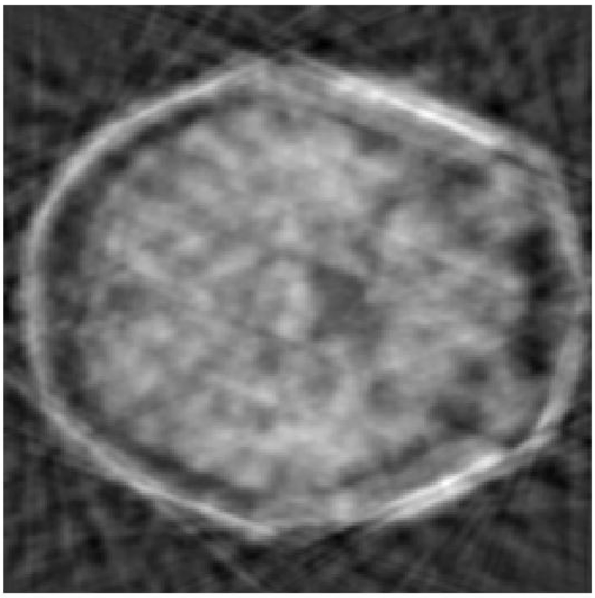

Figure 4: Coupled CS-based Reconstruction using 2 slices, Image 2

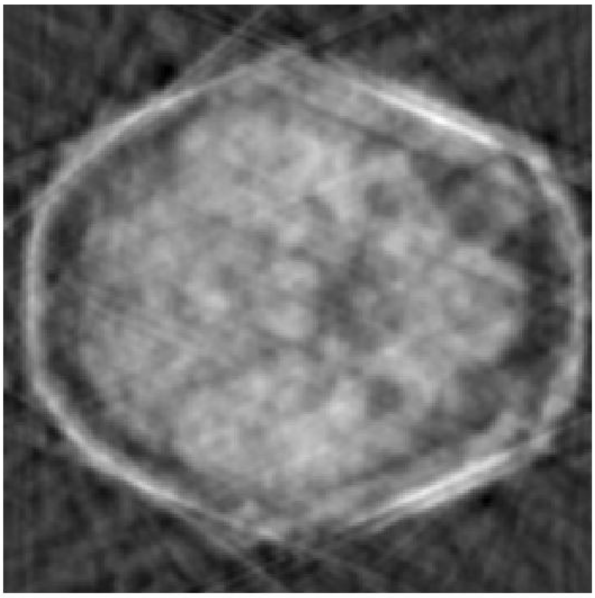

Figure 5: Coupled CS-based Reconstruction using 3 slices, Image 1

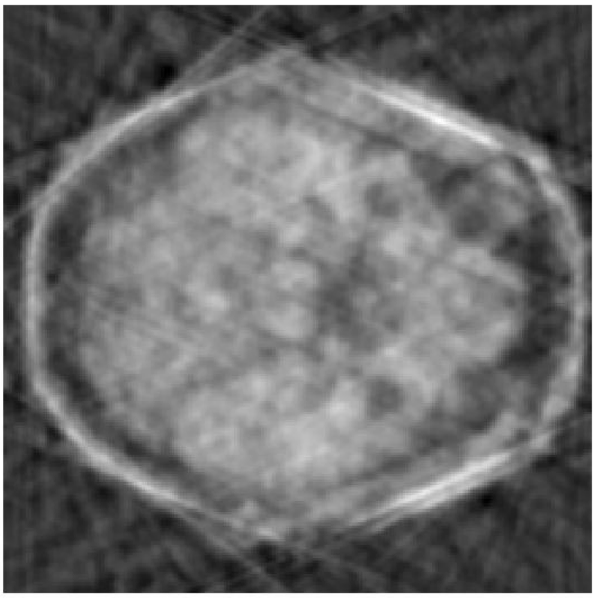

Figure 6: Coupled CS-based Reconstruction using 3 slices, Image 2

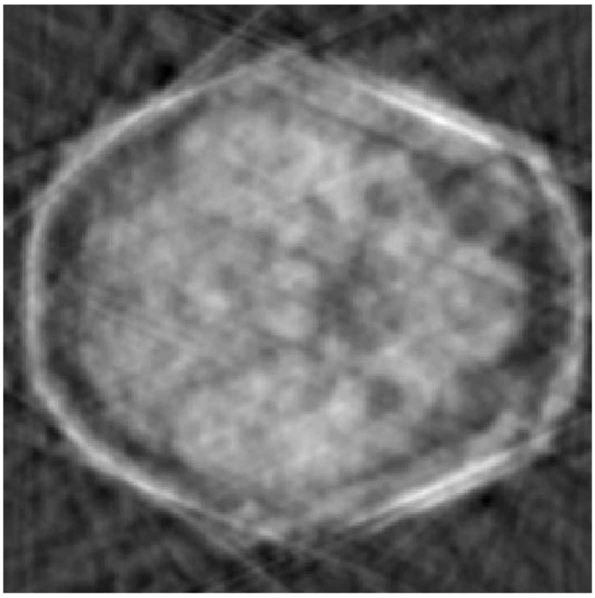

Figure 7: Coupled CS-based Reconstruction using 3 slices, Image 3

Following is the objective function for the case of three similar slices.

$$\begin{aligned}E(\beta_1, \beta_2, \beta_3) &= \|y_1 - R_1 U\beta_1\|^2 + \|y_2 - R_2 U\beta_2\|^2 + \|y_3 - R_3 U\beta_3\|^2 \\ &+ \lambda\|\beta_1\| + \lambda\|\beta_2 - \beta_1\| + \lambda\|\beta_3 - \beta_2\| \\ &= \|y_1 - R_1 U\beta_1\|^2 + \|y_2 - R_2 U(\beta_1 + \Delta\beta_{21})\|^2 + \|y_3 - R_3 U(\beta_1 + \Delta\beta_{21} + \Delta\beta_{32})\|^2 \\ &+ \lambda\|\beta_1\| + \lambda\|\Delta\beta_{21}\| + \lambda\|\Delta\beta_{32}\| \\ &= \left\| \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} - \begin{pmatrix} R_1 U & 0 & 0 \\ R_2 U & R_2 U & 0 \\ R_3 U & R_3 U & R_3 U \end{pmatrix} \begin{pmatrix} \beta_1 \\ \Delta\beta_{21} \\ \Delta\beta_{32} \end{pmatrix} \right\|^2 + \lambda \left\| \begin{pmatrix} \beta_1 \\ \Delta\beta_{21} \\ \Delta\beta_{32} \end{pmatrix} \right\|\end{aligned} \tag{1}$$

Here $x_1 = U\beta_1$, $x_2 = U\beta_2$ and $x_3 = U\beta_3$ represent two consecutive slices of brain where each slice is a 2D image. y_1, y_2 and y_3 represent their tomographic projections expressed as 1D vectors. R_1, R_2 and R_3 denote the Radon-based forward models for different angle sets.
Also $\Delta\beta_{21} = \beta_2 - \beta_1$ and $\Delta\beta_{32} = \beta_3 - \beta_2$.

Question 3

a) Let $f(x,y) = g(x-x_0, y-y_0)$ Let $x' = x - x_0$ and $y' = y - y0$. Then,

$$R(g(x-x_0,y-y_0))(\rho,\theta) = R(f(x,y))(\rho,\theta) \implies R(g(x-x_0,y-y_0)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\delta(\rho - xcos\theta - ysin\theta)\,dx\,dy$$

$$\implies R(g(x-x_0,y-y_0)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x-x_0,y-y_0)\delta(\rho - xcos\theta - ysin\theta)\,dx\,dy$$

Now let $x_1 = x - x_0$ and $y_1 = y - y_0$. Then, replacing x_1 and y_1 in above gives,

$$R(g(x-x_0,y-y_0)) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1,y_1)\delta(\rho - x_1cos\theta - y_1sin\theta - x_0cos\theta - y_0sin\theta)\,dx_1\,dy_1 \tag{2}$$

Now in (1) ,

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x_1,y_1)\delta(\rho - x_1cos\theta - y_1sin\theta - x_0cos\theta - y_0sin\theta)\,dx_1\,dy_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\delta(\rho - xcos\theta - ysin\theta - x_0cos\theta - y_0sin\theta)\,d$$

as we can just relabel the variables inside of the integrals.
Now,

$$R(g(x,y))(\rho - x_0cos\theta - y_0sin\theta, \theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\delta(\rho - xcos\theta - ysin\theta - x_0cos\theta - y_0sin\theta)\,dx\,dy$$

Hence $R(g(x-x_0,y-y_0))(\rho,\theta) = R(g(x,y))(\rho - x_0cos\theta - y_0sin\theta, \theta)$.

b)for polar coordinates (r,ψ),

$$R(g'(r,\psi))(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g'(r,\psi)\delta(\rho - rcos\psi cos\theta - rsin\psi sin\theta)r\,dr\,d\psi \implies R(g'(r,\psi))(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi-\psi_0)\delta(\rho$$

$$\implies R(g'(r,\psi))(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi-\psi_0)\delta(\rho - r(cos\psi cos\theta + sin\psi sin\theta))r\,dr\,d\psi$$

$$\implies R(g'(r,\psi))(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi-\psi_0)\delta(\rho - rcos(\psi-\theta))r\,dr\,d\psi$$

Now relabelling $\psi = \psi - \psi_0$ in the integral above,

$$\implies R(g'(r,\psi))(\rho,\theta) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi)\delta(\rho - rcos(\psi+\psi_0-\theta))r\,dr\,d\psi$$

Also,

$$R(g(r,\psi))(\rho,\theta-\psi_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi)\delta(\rho - rcos\psi cos(\theta-\psi_0) - rsin\psi sin(\theta-\psi_0))r\,dr\,d\psi$$

$$\implies R(g(r,\psi))(\rho,\theta-\psi_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi)\delta(\rho - r(cos\psi cos(\theta-\psi_0) - sin\psi sin(\theta-\psi_0)))r\,dr\,d\psi$$

$$\implies R(g(r,\psi))(\rho,\theta-\psi_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi)\delta(\rho - r(cos(\psi-(\theta-\psi_0))))r\,dr\,d\psi$$

$$\implies R(g(r,\psi))(\rho,\theta-\psi_0) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(r,\psi)\delta(\rho - rcos(\psi-\theta+\psi_0))r\,dr\,d\psi$$

Hence $R(g(r,\psi))(\rho,\theta-\psi_0) = R(g'(r,\psi))(\rho,\theta)$

c) Let $h = f * k$.

$$h(x,y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)k(x-x_1,y-y_1)\,dx_1\,dy_1$$

Then, Radon transform of h is

$$R_\theta(f*k) = R_\theta(h) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)k(x-x_1,y-y_1)\delta(\rho - xcos\theta - ysin\theta)\,dx_1\,dy_1\,dx\,dy$$

$$\implies R_\theta(f*k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(x-x_1,y-y_1)\delta(\rho - xcos\theta - ysin\theta)\,dx\,dy)\,dx_1\,dy_1$$

Now $(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} k(x-x_1,y-y_1)\delta(\rho - xcos\theta - ysin\theta)\,dx\,dy)$ is the Radon transform $R_\theta(k(x-x_1,y-y_1))(\rho,\theta)$ which is same as $R_\theta(k(x,y))(\rho - x_1cos\theta - y_1sin\theta,\theta)$ by Shift property proved above. Hence,

$$R_\theta(f*k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)R_\theta(k(x,y))(\rho - x_1cos\theta - y_1sin\theta,\theta)\,dx_1\,dy_1$$

in the above integral we can insert a new integration of ρ' with Dirac delta function as

$$R_\theta(f*k) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)R_\theta(k(x,y))(\rho-\rho',\theta)\delta(\rho' - x_1cos\theta - y_1sin\theta,\theta)\,dx_1\,dy_1\,d\rho'$$

now $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x_1,y_1)\delta(\rho' - x_1cos\theta - y_1sin\theta)\,dx_1\,dy_1$ is Radon transform $R_\theta f(\rho',\theta)$. Hence,

$$R_\theta(f*k) = \int_{-\infty}^{\infty} R_\theta(f)(\rho',\theta)R_\theta(k)(\rho-\rho',\theta)\,d\rho'$$

This is convolution with respect to the ρ affine parameter.Hence, $R_\theta(f*k) = R_\theta(f) * R_\theta(k)$

Question 4

From RIP, we have

$$(1-\delta s)\|\theta\|^2 \leq \|A\theta\|^2 \leq (1+\delta s)\|\theta\|^2 \tag{3}$$

Let A_Γ denote the subset of A with the support set $|\Gamma| \leq S$ which is a subset of possible nonzero indices of vector θ. Let us denote the maximum and minimum eigenvalues of A_Γ by λ_{max} and λ_{min}. They are as follows

$$\lambda_{max} = max_{\theta_\Gamma \in R^S} \frac{\|A_\Gamma \theta_\Gamma\|^2}{\|\theta_\Gamma\|^2} \tag{4}$$

$$\lambda_{min} = min_{\theta_\Gamma \in R^S} \frac{\|A_\Gamma \theta_\Gamma\|^2}{\|\theta_\Gamma\|^2} \tag{5}$$

Also we know that

$$\delta_s = max\{\lambda_{max} - 1, 1 - \lambda_{min}\} \tag{6}$$

Now consider the matrix $A_s^T A_s$ which is a symmetric matrix. Since it is given in the question that the columns of the matrix A are unit norm. Therefore the diagonal elements of the matrix $A_s^T A_s$ must be equal to 1.

The Gershgorin's disc theorem states that for any square matrix B, for every eigenvalue λ, $\exists i \in \{1, ..., n\}$ such that the following is true

$$B_{ii} - r_i \leq \lambda \leq B_{ii} + r_i \tag{7}$$

where B_{ii} is the diagonal element of the i^{th} row and r_i is the sum of absolute values of the off diagonal elements of the i^{th} row.

Now $A_s^T A_s$ is a square matrix so the equation 7 holds for it. But since A has unit norm columns as given in the question, hence $B_{ii} = 1$ in this case.

Now the mutual coherence for the matrix A is defined as follows.

$$\mu(A) = max_{i,j,i\neq j} \frac{|A_i^t A_j|}{\|A_i\|_2 \|A_j\|_2} \tag{8}$$

But since the columns of A have unit norm, the above reduces to

$$\mu(A) = max_{i,j,i\neq j} |A_i^t A_j| \tag{9}$$

In particular, μ is the maximum value of any off diagonal element of the matrix $A_s^T A_s$ and since there are $s-1$ off diagonal elements in any row of $A_s^T A_s$, therefore we have $r_i \leq \mu(s-1)$. Negation of this also gives $-\mu(s-1) \leq -r_i$ and the equation 7 reduces to

$$1 - \mu(s-1) \leq \lambda \leq 1 + \mu(s-1) \tag{10}$$

This can be further broken down to

$$1 - \mu(s-1) \leq \lambda_{min} \leq \lambda_{max} \leq 1 + \mu(s-1) \tag{11}$$

Now if $\delta_s = \lambda_{max} - 1$ then $1 + \delta_s \leq 1 + \mu(s-1)$ which means $\delta_s \leq \mu(s-1)$ and otherwise if $\delta_s = 1 - \lambda_{min}$ then $1 - \mu(s-1) \leq 1 - \delta_s$ which also means $\delta_s \leq \mu(s-1)$. Hence proved.

Question 5

Title: On the use of sensitivity tests in seismic tomography
Author: N. Rawlinson, W. Spakman
Venue & Year Geophysical Journal International, Volume 205, Issue 2, 01 May 2016, Pages 1221–1243
Link: https://doi.org/10.1093/gji/ggw084

The linearized forward equation is

$$d = Gm + \epsilon \tag{12}$$

where $\epsilon = \epsilon_t + \epsilon_d + \epsilon_p + \epsilon_l$ comprises all sources of observation error and errors introduced by the various assumptions as stated below.

ϵ_t represents the discrepancy between d_E and its theoretical prediction $g(m_E)$ and the magnitude of ϵ_t is a function of the approximations made in solving the forward problem. ϵ_d is the observational noise. ϵ_p is the implied parameterization error. ϵ_l is the linearization error.

The general nonlinear forward problem of noise-free data is

$$d_E = g(m_E) + \epsilon_t \tag{13}$$

where m_E represents the distribution of some true-Earth seismic property and g is a nonlinear integral operator. After adding observational noise it becomes

$$d = g(m_E) + \epsilon_t + \epsilon_d \tag{14}$$

A suitable model parameterization(with some error) is assumed that projects m_E on the model vector m_p and the matrix representation of the integral equations as a local linear representation of $g(m_p)$ about the true model m_p is defined. This gives us

$$d = G_p m_p + \epsilon_t + \epsilon_d + \epsilon_p \tag{15}$$

where G_p is the observation matrix relating the data d_p to the true Earth model m.

At this point, the above is still a nonlinear relationship and one assumes an approximation G which is a known matrix obtained from integration over a background reference model.

$$G_p m_p = (G - E)m_p = Gm_p + \epsilon_l \tag{16}$$

where $E = G - G_p$ is the linearization error in the observation matrix.

Despite all this, we cannot expect to retrieve m_p from the inversion of $d = Gm_p + \epsilon$, as it is only one of many data satisfying solutions. To make this explicit, m_p is replaced by m, which denotes any model that can fit the data equally as written in the first equation.

The cost function associated with the linearized and regularized least squares inversion is

$$S(m) = (d - Gm)^T C_d^{-1}(d - Gm) + \alpha^2 m^T D^T Dm \tag{17}$$

The first term is the quadratic data misfit scaled by the prior data covariance C_d and the second term defines the penalties on model attributes through a predesigned damping matrix αD where α is the tuning parameter.