College of Engineering
Daniel Guggenheim School of Aerospace Engineering

ISYE 7750: Mathematical Foundations of Machine Learning

Homework 6

Professor:
Ashwin Pananjady
Gtech ISYE Professor

Student:
Tomoki Koike
AE MS Student

November 13, 2022

Table of Contents

I Problem One

Suppose that two random variables (X, Y) have joint pdf $f_{X,Y}(x, y)$. Find an expression for the pdf $f_Z(z)$ where $Z = X + Y$. You can start by realizing that

$$F_Z(u \mid X = \beta) = \mathbb{P}[Z \leq u \mid X = \beta] = \mathbb{P}[Y \leq u - \beta \mid X = \beta].$$

You can combine the expressions above by integrating over β, and see that the resulting expression corresponds to an integral of $f_{X,Y}(x, y)$ over a half plane. From this, you can get the pdf for Z by applying the Fundamental Theorem of Calculus. How does your expression simplify if X and Y are independent? (Convolution!)

Solution

From the given hint we can proceed to solve

$$\begin{aligned} F_Z(u \mid X = \beta) &= \mathbb{P}[Z \leq u \mid X = \beta] = \mathbb{P}[Y \leq u - \beta \mid X = \beta] \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{u-\beta} f_{XY}(\beta, y) dy d\beta. \end{aligned}$$

Now if we apply the Fundamental Theorem of Calculus we have

$$\begin{aligned} f_Z(u) &= \frac{\mathrm{d}}{\mathrm{d}u} F_Z(u \mid X = \beta) \\ &= \int_{-\infty}^{\infty} \left[\frac{\mathrm{d}}{\mathrm{d}u} \int_{-\infty}^{u-\beta} f_{XY}(\beta, y) dy \right] d\beta \\ &= \int_{-\infty}^{\infty} f_{XY}(\beta, u - \beta) d\beta. \end{aligned}$$

Thus, we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_{XY}(x, z - x) dx. \tag{I.1}$$

If X and Y are independent, then we have

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx. \tag{I.2}$$

II Problem Two

Let $X_1, X_2, \ldots$ be independent uniform random variables,

$$X_n \sim \text{Uniform}(-1/2, 1/2), \quad \text{meaning} \quad f_X(x) = \begin{cases} 1, & -1/2 \le x \le 1/2 \\ 0, & \text{otherwise.} \end{cases}$$

(a) What is the density function for $Y = X_1 + X_2 + X_3$? (If you compute this correctly, you will meet an old friend.)

(b) The *moment generating function* of a random variable is

$$\varphi_X(t) = \mathbb{E}\left[e^{tX}\right].$$

It is a fact that if $\varphi_X(t) = \varphi_W(t)$ for all t, then X and W have the same distribution. It is a fact that if $G \sim \mathcal{N}(0, \sigma^2)$, then $\varphi_G(t) = e^{\sigma^2 t^2/2}$. Let

$$Y_N = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} X_n.$$

Find an expression for $\varphi_{Y_N}(t)$. Plot $\varphi_{Y_N}(t)$ and $\varphi_G(t)$ for $\sigma^2 = \text{Var}[Y] = \text{Var}[X_n] = 1/12$ on the same set of axes for $N = 1, 2, 5, 10$ and $0 \le t \le 5$. What might you conclude about Y_N as $N \to \infty$? (***Bonus question**: argue rigorously that $\varphi_{Y_N}(t) \to \varphi_G(t)$ for all t.*)

(c) It is a fact that if $\phi(z)$ is a monotonically increasing function, then for any random variable Z,

$$\mathbb{P}[Z > u] = \mathbb{P}[\phi(Z) > \phi(u)].$$

Use $\phi(z) = e^{tz}$ and the Markov inequality to derive a bound on $\mathbb{P}[Z_N > u]$, where

$$Z_N = \frac{1}{N} \sum_{n=1}^{N} X_n.$$

For the special case of $t = 4u/N$, compare this bound, as a function of u, to that obtained using the Chebyshev inequality.

Solution

Question (a)

For this question, we will use the convolution property that we derived in the previous question. Let, $Z = X_1 + X_2$, then $Y = X_3 + Z$. Performing a 2 step convolution we are able to obtain the pdf of Y. For the first step,

$$f_Z(z) = f_{X_1}(z) * f_{X_2}(z) = \int_{-\infty}^{\infty} f_{X_1}(x) f_{X_2}(z - x) dx. \tag{II.1}$$

Now, for the integrand to be 1, the domain should be $D = \{x \mid x \in [-1/2, 1/2] \cap [z - 1/2, z + 1/2]\}$ where the range is $z \in [-1, 1]$. Thus,

$$f_Z(z) = \begin{cases} \int_{-1/2}^{z+1/2} 1 dx = z + 1 & -1 \le z \le 0 \\ \int_{z-1/2}^{1/2} 1 dx = 1 - z & 0 \le z \le 1 \\ 0 & \text{otherwise} \end{cases} \tag{II.2}$$

Then for the second convolution we have

$$f_Y(y) = f_Z(y) * f_{X_3}(y) = \int_{-\infty}^{\infty} f_Z(y-z) f_{X_3}(z) dz = \int_{-1/2}^{1/2} f_Z(y-z) dz = \int_{y-1/2}^{y+1/2} f_{X_3}(z) dz. \tag{II.3}$$

Thus, we have

$$f_Y(y) = \begin{cases} \int_{-1}^{y+1/2} (z+1) dz = \frac{y^2}{2} + \frac{3y}{2} + \frac{9}{8} & -3/2 \le y \le -1/2 \\ \int_{y-1/2}^{0} (z+1) dz + \int_{0}^{y+1/2} (1-z) dz = -y^2 + \frac{3}{4} & -1/2 \le y \le 1/2 \\ \int_{y-1/2}^{1} (1-z) dz = \frac{y^2}{2} - \frac{3y}{2} + \frac{9}{8} & 1/2 \le y \le 3/2 \\ 0 & \text{otherwise} \end{cases} \tag{II.4}$$

Question (b)

Firstly, the pdf for a uniform distribution of $X_n \sim \mathcal{U}[-1/2, 1/2]$ is

$$\begin{aligned} \phi_{X_N}(t) &= \int_{-\infty}^{-1/2} 0 \cdot e^{tx} dx + \int_{-1/2}^{1/2} e^{tx} dx + \int_{1/2}^{\infty} 0 e^{tx} dx \\ &= \left[\frac{e^{tx}}{t} \right]_{-1/2}^{1/2} = \frac{1}{t} \left(e^{\frac{t}{2}} - e^{-\frac{t}{2}} \right). \end{aligned} \tag{II.5}$$

Keeping in mind that all X_i are independent from each other, the moment generating function of Y_N becomes

$$\begin{aligned} \phi_{Y_N} &= \mathbb{E}\left[e^{tY_N}\right] = \mathbb{E}\left[e^{\frac{t}{\sqrt{N}}(X_1+X_2+\cdots+X_N)}\right] \\ &= \mathbb{E}\left[e^{\frac{t}{\sqrt{N}}X_1}\right] \mathbb{E}\left[e^{\frac{t}{\sqrt{N}}X_2}\right] \cdots \mathbb{E}\left[e^{\frac{t}{\sqrt{N}}X_N}\right] \\ &= \left[\frac{\sqrt{N}\left(e^{\frac{t}{2\sqrt{N}}} - e^{-\frac{t}{2\sqrt{N}}}\right)}{t} \right] \left[\frac{\sqrt{N}\left(e^{\frac{t}{2\sqrt{N}}} - e^{-\frac{t}{2\sqrt{N}}}\right)}{t} \right] \cdots \left[\frac{\sqrt{N}\left(e^{\frac{t}{2\sqrt{N}}} - e^{-\frac{t}{2\sqrt{N}}}\right)}{t} \right] \\ &= \left[\frac{\sqrt{N}\left(e^{\frac{t}{2\sqrt{N}}} - e^{-\frac{t}{2\sqrt{N}}}\right)}{t} \right]^N \end{aligned} \tag{II.6}$$

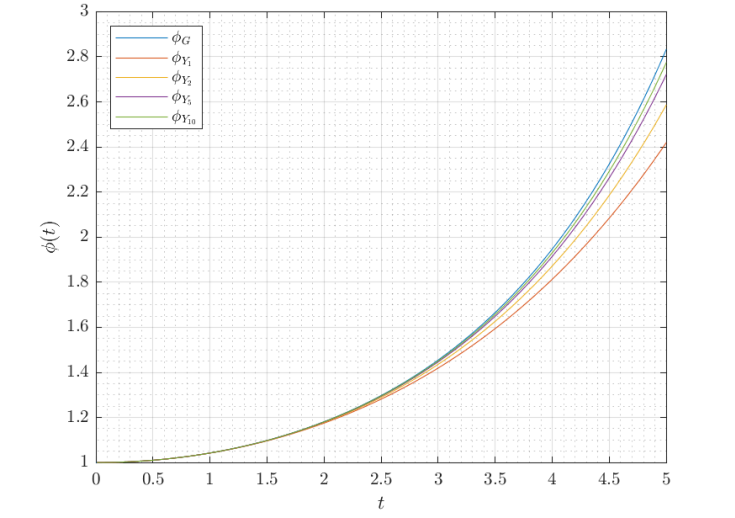


Figure 1: ϕ_{Y_N} for N=1,2,5,10 with ϕ_G.

From this figure we can see that as $N \to \infty$ the distribution Y_N converges to G. Now let us prove this rigorously. First we can rewrite the moment generating function of (II.6) as

$$\phi_{Y_N} = \left[\frac{\sinh(t/2\sqrt{N})}{t/2\sqrt{N}}\right]^N \tag{II.7}$$

Now using the Taylor expansion we have

$$\begin{aligned}
\phi_{Y_N} &= \left[\frac{\frac{t}{2\sqrt{N}} + \frac{1}{3!}\left(\frac{t}{2\sqrt{N}}\right)^3 + \frac{1}{5!}\left(\frac{t}{2\sqrt{N}}\right)^5 + \cdots}{\frac{t}{2\sqrt{N}}}\right]^N \\
&= \left[1 + \frac{1}{3!}\left(\frac{t}{2\sqrt{N}}\right)^2 + \frac{1}{5!}\left(\frac{t}{2\sqrt{N}}\right)^4 + \cdots\right]^N \\
&\approx \left[1 + \frac{1}{3!}\left(\frac{t}{2\sqrt{N}}\right)^2\right]^N \\
&= \left[1 + \frac{1}{6}\left(\frac{t}{2\sqrt{N}}\right)^2\right]^N
\end{aligned}$$

Then

$$\lim_{N\to\infty} \phi_{Y_N} \approx \lim_{N\to\infty} \left[1 + \frac{1}{6}\left(\frac{t}{2\sqrt{N}}\right)^2\right]^N$$

Where if you use the relation

$$\lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n = e^x,$$

you have

$$\lim_{N\to\infty} \phi_{Y_N} \approx \lim_{N\to\infty} \left[1 + \frac{1}{6}\left(\frac{t}{2\sqrt{N}}\right)^2\right]^N = \lim_{N\to\infty} \left[1 + \frac{t^2}{24N}\right]^N = e^{\frac{t^2}{24}} = \phi_G(t).$$

Hence, for $N \to \infty$ we have $\phi_{Y_N} \to \phi_G$.

Question (c)

Solving the probability we have using Markov inequality

$$\mathbb{P}\left[Z_N > u\right] = \mathbb{P}\left[\phi(Z_N) > \phi(u)\right] \leq \frac{\mathbb{E}\left[\phi(Z_N)\right]}{e^{tu}}.$$

Similar to the computation in the previous question we have

$$\mathbb{E}\left[e^{tZ_N}\right] = \left(\frac{e^{\frac{t}{2N}} - e^{-\frac{t}{2N}}}{t/N}\right)^N$$

Thus,

$$\mathbb{P}\left[Z_N > u\right] \leq e^{-tu}\left(\frac{e^{\frac{t}{2N}} - e^{-\frac{t}{2N}}}{t/N}\right)^N \xrightarrow{t\to\frac{4u}{N}} e^{-\frac{4u^2}{N}}\left(\frac{e^{\frac{2u}{N^2}} - e^{-\frac{2u}{n^2}}}{4u/N^2}\right)^N$$

We know that Z_N represents a sample mean of the independent uniform random variables. Therefore, the mean for Z_N is equal to the mean of X_i, and let this be $\mu = 0$ for X_i. And the Variance of the sample mean is σ^2/N where $\sigma^2 = (1/2 + 1/2)^2/12 = 1/12$. Hence, from Chebyshev inequality we know that

$$\mathbb{P}\left[|Z_N - \mu| > u\right] = \mathbb{P}\left[Z_N > u\right] \leq \frac{\sigma^2}{Nu^2} = \frac{1}{12Nu^2}.$$

If we plot this out for different N and u values we have the following

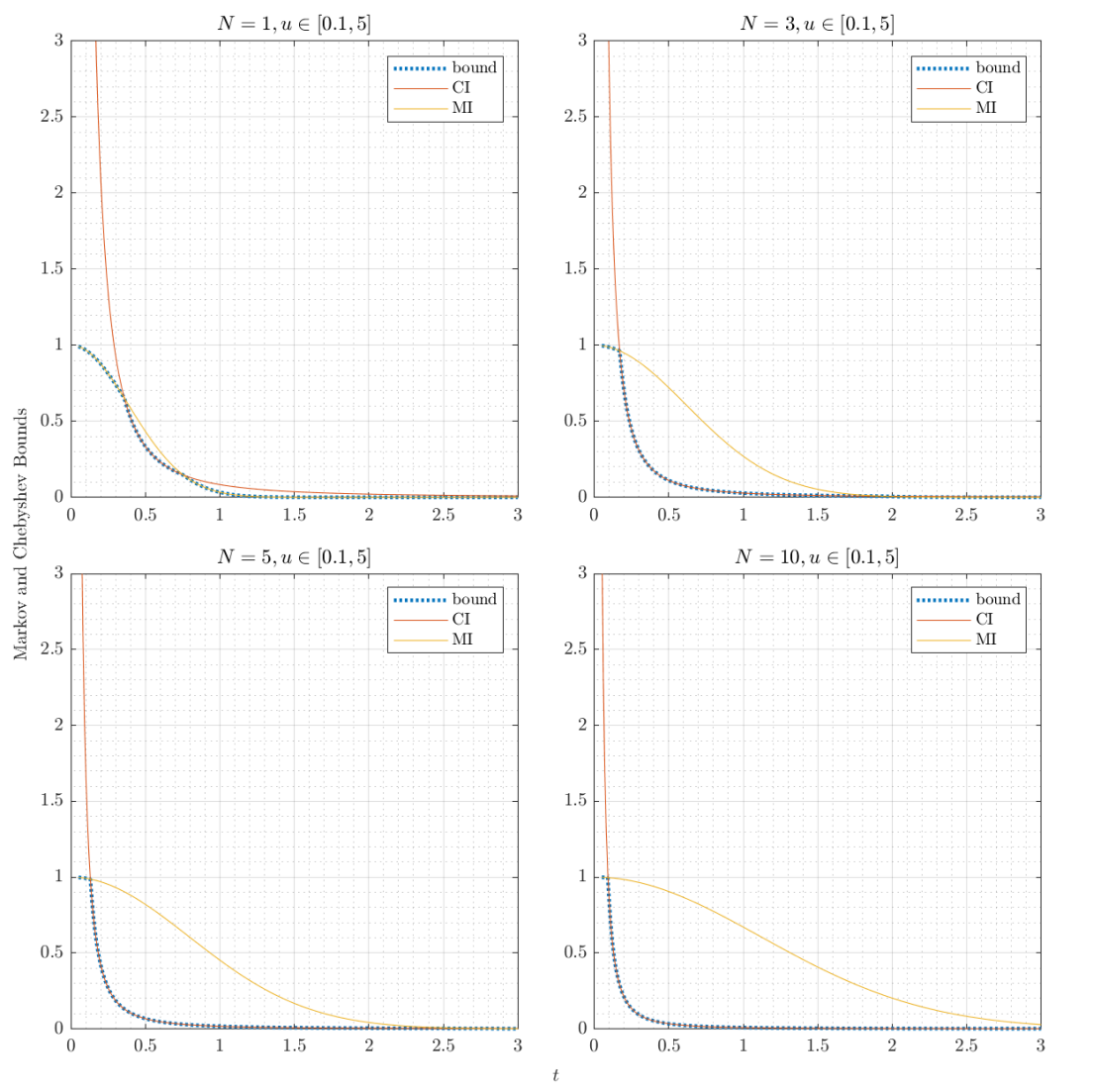


Figure 2: The Markov inequality and Chebyshev inequality bounds for different N and u values.

This plot shows how for different u and N values the alter the lowest bound between the Markov and Chebyshev inequality. Particularly, the dotted curve on the plot above is the actual lowest bound of the two and for low N values we see that there is 2 instances where it switches between the two; however, as N increases there is only 1 switching. For all of them we see that the Markov inequality is the lower bound for very small u values since the bound from Chebyshev diverges for when u approaches 0.

III Problem Three

Let $Z_1, \dots, Z_N$ be a sequence of independent Gaussian random variables with mean 0 and variance 1. You observe the random vector X in $\mathbb{R}^N$ that is generated through the autoregressive process

$$X_k = \begin{cases} Z_1, & k = 1 \\ aX_{k-1} + Z_k, & k > 1. \end{cases}$$

Given $X = \mathbf{x}$, find the MLE for $a \in \mathbb{R}$. (Hint: Conditional independence.) (Further hint: The conditional independence structure makes this a Markov process, meaning that we can factor the distribution for $X \in \mathbb{R}^N$ as

$$f_X(\mathbf{x}) = f_{X_1}(x_1) f_{X_2}(x_2|x_1) f_{X_3}(x_3|x_2) \cdots f_{X_N}(x_N|x_{N-1}).$$

Solution

Let the parameter found for the maximum likelihood be a. Since we know from the definition of the conditional density

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_X(x)}, \tag{III.1}$$

we can say

$$\begin{aligned} f_a(X_1, \dots, X_N) &= f_a(X_N|X_{N-1}, \dots, X_1) f_a(X_1, \dots, X_{N-1}) \\ &= f_a(X_N|X_{N-1}, \dots, X_1) f_a(X_{N-1}|X_{N-2}, \dots, X_1) f_a(X_1, \dots, X_{N-2}) \\ &\vdots \\ &= \prod_{k=1}^{N} f_a(X_k|X_{k-1}, \dots, X_1) f_a(X_1) \end{aligned} \tag{III.2}$$

Now, from conditional distribution, we have $X_k|X_{k-1}, \dots, Z_1 \triangleq X_k|X_{k-1}$, then it follows that

$$f_a(X_k|X_{k-1}, \dots, X_1) = f_a(X_k|X_{k-1}),$$

and

$$f_a(X_1, \dots, X_N) = \prod_{k=1}^{N} f_a(X_k|X_{k-1}) f_a(X_1)$$

Then,

$$X_k|X_{k-1} \sim \mathcal{N}\left(\mathbb{E}\left[X_k|X_{k-1}\right], \mathrm{Var}\left[X_k|X_{k-1}\right]\right).$$

Since from the structure of the autoregressive process we know that

$$\mathbb{E}\left[X_k|X_{k-1}\right] = aX_{k-1}, \qquad \mathrm{Var}\left[X_k|X_{k-1}\right] = \mathrm{Var}\left[Z_k\right] = 1.$$

Now the conditional density becomes

$$\begin{aligned} f_a(X_k|X_{k-1}) &= \frac{1}{\sqrt{2\pi \mathrm{Var}\left[X_k|X_{k-1}\right]}} \exp\left(-\frac{(X_k - \mathbb{E}\left[X_k|X_{k-1}\right])^2}{2\mathrm{Var}\left[X_k|X_{k-1}\right]}\right) \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_k - aX_{k-1})^2}{2}\right). \end{aligned} \tag{III.3}$$

The conditional likelihood is

$$\begin{aligned} L(a) &= f_a(X_2, \ldots, X_N | X_1) = \prod_{k=1}^{N} f_a(X_k | X_{k-1}) \\ &= \prod_{k=1}^{N} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_k - aX_{k-1})^2}{2} \right) \\ &= (2\pi)^{-\frac{N}{2}} \exp\left[-\frac{X_1^2 + \sum_{n=2}^{N} (X_n - aX_{n-1})^2}{2} \right] \end{aligned}$$

if we take the natural log of this for the log likelihood

$$\begin{aligned} \ell(a) &= \ln L(a) = \sum_{k=1}^{N} \ln f_a(X_k | X_{k-1}) \\ &= -\frac{N}{2} \ln(2\pi) - \frac{1}{2} \left[-\frac{X_1^2 + \sum_{n=2}^{N} (X_n - aX_{n-1})^2}{2} \right] \end{aligned}$$

Taking the derivative of the log likelihood function w.r.t a we have

$$\ell'(a|X) = \sum_{n=2}^{N} X_{n-1}(X_n - aX_{n-1})$$

by settings this to 0 we have

$$\begin{aligned} \sum_{n=2}^{N} X_n X_{n-1} - \sum_{n=2}^{N} a X_{n-1}^2 &= 0 \\ a &= \frac{\sum_{n=2}^{N} X_n X_{n-1}}{\sum_{n=2}^{N} X_{n-1}^2} \end{aligned}$$

Thus, the MLE becomes

$$\hat{a}_{mle} = \frac{\sum_{n=2}^{N} X_n X_{n-1}}{\sum_{n=2}^{N} X_{n-1}^2} \tag{III.4}$$

IV Problem Four

Let X be a Gaussian random vector taking values in $\mathbb{R}^N$, let E be a Gaussian random vector taking values in $\mathbb{R}^M$, and let $\mathbf{A}$ be a $M \times N$ matrix. We have

$$X \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_x), \quad E \sim \mathcal{N}(\mathbf{0}, \mathbf{R}_e), \quad X, E \text{ independent.}$$

We will make observation of the random vector

$$Y = \mathbf{A}X + E.$$

(a) From the lecture notes, it is clear that Y is a Gaussian random vector in $\mathbb{R}^M$ and that $\mathbb{E}[Y] = \mathbf{0}$. Find the covariance matrix for the Gaussian random vector $\begin{bmatrix} X & Y \end{bmatrix}^\top$ that takes values in $\mathbb{R}^{N+M}$.

(b) Suppose we observe $Y = \mathbf{y}$. What is the minimum mean-square error estimate of X given $Y = \mathbf{y}$?

(c) Suppose $\mathbf{R}_x = \sigma_x^2 \mathbf{I}$ and $\mathbf{R}_e = \sigma_e^2 \mathbf{I}$. In this case, your MMSE estimator should look familiar, and you should see immediately that $\hat{x}_{MMSE}$ is in the row space of $\mathbf{A}$. What are the $\hat{\alpha}_n$ is the expression below?

$$\hat{\mathbf{x}}_{MMSE} = \sum_{n=1}^{N} \alpha_n \mathbf{v}_n, \quad \text{where the } \mathbf{v}_n \text{ are the right singular vectors of } \mathbf{A}.$$

(d) Take $\mathbf{R}_x$ and $\mathbf{R}_e$ as in part (c), and assume that $\mathbf{A}$ has full column rank. What is MSE $\mathbb{E}\left[\|\hat{\mathbf{x}}_{MMSE} - X\|_2^2\right]$ of the MMSE estimate $\hat{\mathbf{x}}_{MMSE}$?

Solution

Question (a)

Let $\mathbf{Z} = \begin{bmatrix} X & Y \end{bmatrix}^\top$, then the covariance of this becomes

$$\begin{aligned} \mathbf{R}_z &= \mathbb{E}\left[(\mathbf{Z} - \mu_Z)(\mathbf{Z} - \mu_Z)^\top\right] = \mathbb{E}\left[\mathbf{Z}\mathbf{Z}^\top\right] \\ &= \mathbb{E}\left[\begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} \begin{bmatrix} \mathbf{X}^\top & \mathbf{Y}^\top \end{bmatrix}\right] = \mathbb{E}\left[\begin{bmatrix} \mathbf{X}\mathbf{X}^\top & \mathbf{X}\mathbf{Y}^\top \\ \mathbf{Y}\mathbf{X}^\top & \mathbf{Y}\mathbf{Y}^\top \end{bmatrix}\right] \\ &= \begin{bmatrix} \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right] & \mathbb{E}\left[\mathbf{X}\mathbf{Y}^\top\right] \\ \mathbb{E}\left[\mathbf{Y}\mathbf{X}^\top\right] & \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top\right] \end{bmatrix} \end{aligned} \tag{IV.1}$$

Now since, we know that $\mathbb{E}\left[\mathbf{X}\mathbf{X}^\mathbf{T}\right] = \mathbf{R}_x$, we have to compute the others in the matrix above. Thus,

$$\mathbb{E}\left[\mathbf{X}\mathbf{Y}^\top\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top + \mathbf{X}\mathbf{E}^\top\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^\top\right]\mathbf{A}^\top + \mathbb{E}[X]\,\mathbb{E}\left[E^\top\right] = \mathbf{R}_x\mathbf{A}^\top \tag{IV.2}$$

and

$$\begin{aligned} \mathbb{E}\left[\mathbf{Y}\mathbf{Y}^\top\right] &= \mathbb{E}\left[\mathbf{A}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top + \mathbf{A}\mathbf{X}\mathbf{E}^\top + \mathbf{E}\mathbf{X}^\top\mathbf{A}^\top + \mathbf{E}\mathbf{E}^\top\right] \\ &= \mathbf{A}\mathbb{E}\left[X X^\top\right]\mathbf{A}^\top + \mathbb{E}\left[\mathbf{E}\mathbf{E}^\top\right] = \mathbf{A}\mathbf{R}_x\mathbf{A}^\top + \mathbf{R}_e \end{aligned} \tag{IV.3}$$

Thus,

$$\mathbf{R}_z = \begin{bmatrix} \mathbf{R}_x & \mathbf{R}_{xy} \\ \mathbf{R}_{xy}^\top & \mathbf{R}_y \end{bmatrix} = \begin{bmatrix} \mathbf{R}_x & \mathbf{R}_x\mathbf{A}^\top \\ \mathbf{A}\mathbf{R}_x & \mathbf{A}\mathbf{R}_x\mathbf{A}^\top + \mathbf{R}_e \end{bmatrix}. \tag{IV.4}$$

Question (b)

We want to find

$$\underset{g}{\operatorname{argmin}}\, \mathbb{E}\left[(\mathbf{X}-\mathbf{g})^2|\mathbf{Y}=\mathbf{y}\right] \tag{IV.5}$$

which has a solution of

$$\hat{\mathbf{g}} = \mathbb{E}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right].$$

For this solution the minimum mean-square error becomes

$$\mathbb{E}\left[(\mathbf{X}-\hat{\mathbf{g}})^2|\mathbf{Y}=\mathbf{y}\right] = \mathrm{Var}\left[\mathbf{X}|\mathbf{Y}=\mathbf{y}\right] = \mathbf{R}_{x|y}$$

From Schur's compliment we know the conditional variance is

$$\mathbf{R}_{x|y} = \mathbf{R}_x - \mathbf{R}_{xy}\mathbf{R}_y^{-1}\mathbf{R}_{xy}^\top = \mathbf{R}_x - \mathbf{R}_x\mathbf{A}^\top(\mathbf{A}\mathbf{R}_x\mathbf{A}^\top + \mathbf{R}_e)^{-1}\mathbf{A}\mathbf{R}_x. \tag{IV.6}$$

Question (c)

We know that

$$\begin{aligned}\hat{\mathbf{x}}_{MMSE} = \hat{g} &= \boldsymbol{\mu}_x + \mathbf{R}_x\mathbf{A}^\top(\mathbf{A}\mathbf{R}_x\mathbf{A}^\top + \mathbf{R}_e)^{-1}(\mathbf{y}-\boldsymbol{\mu}_y) = \sigma_x^2\mathbf{A}^\top(\sigma_x^2\mathbf{A}\mathbf{A}^\top + \sigma_e^2\mathbf{I})^{-1}\mathbf{y} \\ &= \mathbf{A}^\top\left(\mathbf{A}\mathbf{A}^\top + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{y} = \left(\mathbf{A}^\top\mathbf{A} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{A}^\top\mathbf{y}\end{aligned}$$

This looks very similar to the regularized least square solution. Thus, with SVD where $\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top$ we can find the singular values σ_n of $\mathbf{A}$ and the left singular vectors to be $\mathbf{u}_n$ then

$$\begin{aligned}\hat{\mathbf{x}}_{MMSE} &= \left(\mathbf{A}^\top\mathbf{A} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{A}^\top\mathbf{y} \\ &= \left(\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{V}\mathbf{V}^\top\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} \\ &= \left(\mathbf{V}\boldsymbol{\Sigma}^\mathbf{2}\mathbf{V}^\top + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{V}\mathbf{V}^\top\right)^{-1}\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} \\ &= \mathbf{V}\left(\boldsymbol{\Sigma}^2 + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\mathbf{V}^\top\mathbf{V}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{y} \\ &= \underbrace{\mathbf{V}\left(\boldsymbol{\Sigma}^2 + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{\Sigma}}_{\boldsymbol{\Psi}}\mathbf{U}^\top\mathbf{y}\end{aligned}$$

This $\boldsymbol{\Psi}$ matrix is diagonal, and let the singular values of $\mathbf{A}$ in $\boldsymbol{\Sigma}$ be σ, then the diagonal entries are

$$\sigma_\psi = \left(\sigma^2 + \frac{\sigma_e^2}{\sigma_x^2}\right)^{-1}\cdot\sigma = \frac{\sigma\sigma_x^2}{\sigma^2\sigma_x^2 + \sigma_e^2}$$

Hence, the minimum mean square error can be represented as

$$\hat{\mathbf{x}}_{MMSE} = \sum_{n=1}^{N}\sigma_{\psi_n}\langle\mathbf{y},\mathbf{u_n}\rangle\mathbf{v}_n = \sum_{n=1}^{N}\frac{\sigma_n\sigma_x^2}{\sigma_n^2\sigma_x^2+\sigma_e^2}\langle\mathbf{y},\mathbf{u}_n\rangle\mathbf{v}_n. \tag{IV.7}$$

where σ_n is the n-th singular value of matrix $\mathbf{A}$.

Question (d)

We compute the below

$$\begin{aligned}
\mathbb{E}\left[\|\hat{\mathbf{x}}_{MMSE}-\mathbf{X}\|_2^2\right] &= \mathrm{tr}\left(\mathbf{R_x}-\mathbf{R_{xy}}\mathbf{R_y^{-1}}\mathbf{R_{xy}^T}\right)\\
&= \mathrm{tr}\left(\mathbf{R_x}-\mathbf{R_x}\mathbf{A}^\top(\mathbf{A}\mathbf{R_x}\mathbf{A}^\top+\mathbf{R_e})^{\mathbf{-1}}\mathbf{A}\mathbf{R_x}\right)\\
&= \mathrm{tr}\left(\sigma_x\mathbf{I}\right)-\sigma_x\mathrm{tr}\left(\left(\mathbf{A}^\top\mathbf{A}+\frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}A^\top A\right) \qquad (\because \text{Question (c)})\\
&= N\sigma_x^2-\sigma_x^2\mathrm{tr}\left(\left(\mathbf{V}\mathbf{\Sigma^2}\mathbf{V}^\top+\frac{\sigma_e^2}{\sigma_x^2}\mathbf{V}\mathbf{V}^\top\right)^{-1}\mathbf{V}\mathbf{\Sigma^2}\mathbf{V}^\top\right)\\
&= N\sigma_x^2-\sigma_x^2\mathrm{tr}\left(\underbrace{\mathbf{V}\left(\mathbf{\Sigma^2}+\frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}}_{\mathbf{C}}\underbrace{\mathbf{\Sigma^2}\mathbf{V}^\top}_{\mathbf{D}}\right)\\
&= N\sigma_x^2-\sigma_x^2\mathrm{tr}\left(\mathbf{\Sigma}^2\left(\mathbf{\Sigma^2}+\frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\right) \qquad (\because \mathrm{tr}\left(\mathbf{CD}\right)=\mathrm{tr}\left(\mathbf{DC}\right))\\
&= N\sigma_x^2-\sigma_x^2\sum_{n=1}^{N}\frac{\sigma_n^2\sigma_x^2}{\sigma_n^2\sigma_x^2+\sigma_e^2}\\
&= \sigma_x^2\left(N-\sum_{n=1}^{N}\frac{\sigma_n^2\sigma_x^2}{\sigma_n^2\sigma_x^2+\sigma_e^2}\right)
\end{aligned}$$

V Problem Five

Let $\mathbf{A}$ be an $M \times N$ matrix with full column rank. Let E be a Gaussian random vector in $\mathbb{R}^M$ with mean $\mathbf{0}$ and covariance $\mathbf{R}_e$. Suppose we observe

$$Y = \mathbf{A}\boldsymbol{\theta}_0 + E,$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^N$ is unknown.

(a) What is the distribution of Y and how does it depend on $\boldsymbol{\theta}_0$?

(b) Find a closed form expression for the maximum likelihood estimate of $\boldsymbol{\theta}_0$. (In this case, we are working from a single sample of a random vector.)

(c) What is the distribution of the MLE estimator $\hat{\boldsymbol{\Theta}}$? Is $\hat{\boldsymbol{\Theta}}$ unbiased?

(d) What is the MSE of the MLE, $\mathbb{E}\left[\|\hat{\boldsymbol{\Theta}} - \boldsymbol{\theta}_0\|_2^2\right]$?

(e) Compute the Fisher information matrix $\mathbf{J}(\boldsymbol{\theta}_0)$ and verify that the MLE meets the Cramer-Rao lower bound.

(f) Defend the following statement: The MLE is the best unbiased estimator of $\boldsymbol{\theta}_0$.

Solution

Question (a)

We can say that the unknown parameter $\boldsymbol{\theta}_0$ is inside the family of normal distributions, i.e. $\boldsymbol{\theta}_0 \sim \mathcal{N}(\theta_1, \theta_2)$ and the mean and the variance determining the distribution of the Y is dictated by the mean, μ and variance, $\mathbf{R}$ of this unknown parameter. Hence, the sum of two Gaussian random variables are also Gaussian and we can say

$$Y \sim \mathcal{N}(\mathbf{A}\boldsymbol{\theta_0}; \mathbf{R}_e)$$

from the previous problem.

Question (b)

The likelihood of this unknown parameter is

$$L(\boldsymbol{\theta}_0; \mathbf{E}) = f_Y(\mathbf{E}; \boldsymbol{\theta}_0) \tag{V.1}$$

and the maximum likelihood estimation is simply the parameters that maximize the likelihood $L(\boldsymbol{\theta}_0; \cdot)$, which is

$$\hat{\boldsymbol{\theta}}_0 = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, L(\boldsymbol{\theta}; \mathbf{E}).$$

where

$$L(\boldsymbol{\theta}_0; \mathbf{E}) = (2\pi)^{-\frac{M}{2}} \det(\mathbf{R}_e)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)^\top \mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)\right].$$

Then

$$\ell(\boldsymbol{\theta}_0; \mathbf{E}) = -\frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln[\det(\mathbf{R}_e)] - \frac{1}{2}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)^\top \mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)$$

Thus, we know that we want to solve the following optimization problem, i.e. the closed form expression for the maximum likelihood estimate

$$\hat{\boldsymbol{\Theta}} = \underset{\boldsymbol{\theta}_0 \in \mathbb{R}^N}{\operatorname{argmin}} \left\|\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)\right\|_2^2 \tag{V.2}$$

Question (c)

If we take the derivative of the log likelihood function we have

$$\ell'(\boldsymbol{\theta}_0, \mathbf{R}_e) = -\frac{1}{2}\left[-2\mathbf{A}^\top\mathbf{R}_e^{-1}9(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)\right] = \mathbf{A}^\top\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0). \tag{V.3}$$

Hence, the MLE is

$$\begin{aligned}
\mathbf{A}^\top\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0) &= 0 \\
\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{y} - \mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A}\boldsymbol{\theta}_0 &= 0 \\
\therefore \hat{\mathbf{\Theta}} &= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{y}.
\end{aligned}$$

The mean is

$$\begin{aligned}
\mathbb{E}\left[\hat{\mathbf{\Theta}}\right] &= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbb{E}\left[\mathbf{y}\right] \\
&= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A}\boldsymbol{\theta}_0 \\
&= \mathbf{A}^{-1}\mathbf{R}_e\mathbf{A}^{-\top}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A}\boldsymbol{\theta}_0 \\
&= \boldsymbol{\theta}_0.
\end{aligned}$$

The variance is

$$\begin{aligned}
\mathrm{Var}\left[\hat{\mathbf{\Theta}}\right] &= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathrm{Var}\left[\mathbf{y}\right]\mathbf{R}_e^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-\top} \\
&= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{R}_e\mathbf{R}_e^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-\top} \\
&= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{R}_e\mathbf{R}_e^{-1}\mathbf{A}(\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-\top} \\
&= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{R}_e\mathbf{R}_e^{-1}\mathbf{A}\mathbf{A}^{-1}\mathbf{R}_e\mathbf{A}^{-\top} \\
&= (\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}.
\end{aligned}$$

Since, the expectation of the MLE is exactly θ_0, it is unbiased.

Question (d)

$$\begin{aligned}
\mathbb{E}\left[\left\|\hat{\mathbf{\Theta}} - \boldsymbol{\theta}_0\right\|_2^2\right] &= \mathrm{tr}\left(\mathrm{Var}\left[\hat{\mathbf{\Theta}}\right]\right) + \left\|\mathbb{E}\left[\hat{\mathbf{\Theta}}\right] - \boldsymbol{\theta}_0\right\|_2^2 \\
&= \mathrm{tr}\left(\mathrm{Var}\left[\hat{\mathbf{\Theta}}\right]\right) \\
&= \mathrm{tr}\left((\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A})^{-1}\right)
\end{aligned}$$

Question (e)

The score function is

$$s(\boldsymbol{\theta}_0; \mathbf{y}) = \nabla_{\theta_0}\ell(\boldsymbol{\theta}_0; \mathbf{y}) = \ell'(\boldsymbol{\theta}_0; \mathbf{y}) = \mathbf{A}^\top\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0).$$

Then it follows

$$ss^\top = \mathbf{A}^\top\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0)^\top\mathbf{R}_e^{-\top}\mathbf{A}.$$

Now since

$$\begin{aligned}
\mathbb{E}\left[s\right] &= \mathbf{A}^\top\mathbf{R}_e^{-1}\mathbb{E}\left[\mathbf{y}\right] - \mathbf{A}^\top\mathbf{R}_e^{-1}(\mathbf{y} - \mathbf{A}\boldsymbol{\theta}_0) = 0 \\
\mathrm{Var}\left[s\right] &= \mathrm{Var}\left[\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{y} - \mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A}\boldsymbol{\theta}_0\right] = \mathrm{Var}\left[\mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{y}\right] = \mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{R}_e(\mathbf{A}^\top\mathbf{R}_e^{-1})^\top = \mathbf{A}^\top\mathbf{R}_e^{-1}\mathbf{A}.
\end{aligned}$$

Hence, $s \sim \mathcal{N}(0, \mathbf{A}^\top \mathbf{R}_e^{-1} \mathbf{A})$. Which allows us to use the following relation,

$$\text{if } \mathbf{z} \sim \mathcal{N}(\mathbf{m}, \mathbf{\Sigma}) \text{ then } \mathbb{E}\left[\mathbf{z}\mathbf{z}^\top\right] = \mathbf{\Sigma} + \mathbf{m}\mathbf{m}^\top.$$

Therefore,

$$J(\hat{\mathbf{\Theta}}) = \mathbb{E}\left[ss^\top\right] = \mathbf{A}^\top \mathbf{R}_e^{-1} \mathbf{A}. \tag{V.4}$$

Now from Question (d) we know that $\text{MSE}(\hat{\mathbf{\Theta}} = \text{tr}\left(()\right)(\mathbf{A}^\top \mathbf{R}_e^{-1} \mathbf{A})^{-1})$. Further,

$$\text{tr}\left(J(\hat{\mathbf{\Theta}})\right) = \text{tr}\left(()\right)(\mathbf{A}^\top \mathbf{R}_e^{-1} \mathbf{A})^{-1})$$

Thus, the Cramer-Rao Lower bound

$$\text{MSE}(\hat{\mathbf{\Theta}}) = \text{tr}\left(()\,\hat{\mathbf{R}}\right) \geq \text{tr}\left(J(\hat{\mathbf{\Theta}})\right) \tag{V.5}$$

holds true.

Question (f)

Since from our course notes, we know that the MLE suffices the three main properties unbias, consistency, and efficiency, which allows an optimal estimator. Additionally, from our analysis from above, we see that the MSE of the MLE is equal to the Cramer-Rao lower bound meaning that achieves the minimum error for an estimation. These facts affirm that the MLE is the best unbiased estimator of θ_0.

VI Problem Six

A Cauchy random variable with "location parameter" ν has a density function

$$f_X(x;\nu) = \frac{1}{\pi(1+(x-\nu)^2)}, \quad x \in \mathbb{R}. \tag{VI.1}$$

Despite its simple definition, this is a strange animal. First of all, its mean is not defined, as the integral $\int x/(1+x^2)\ \underset{\cdot}{\mathrm{x}}$ is not absolutely convergent. It is also easy to see that the variance is infinite. But as you can see (especially if you sketch it), the density is symmetric around ν, and ν is certainly the median.
Let $X_1, X_2, \ldots, X_N$ be iid Cauchy random variables distributed as in (VI.1). From observed data $X_1 = x_1, \ldots, X_N = x_N$, we will compare three estimators: the sample mean

$$\hat{\nu}_{mn} = \frac{1}{N}\sum_{n=1}^{N} x_n,$$

the sample median

$$\hat{\nu}_{md} = \begin{cases} x_{((N+1)/2)}, & \text{N odd,} \\ \frac{x_{(N/2)}+x_{(N/2+1)}}{2}, & \text{N even,} \end{cases}$$

where $x_{(i)}$ is the ith largest value in $\{x_1, \ldots, x_N\}$, and the MLE

$$\hat{\nu}_{mle} = \underset{\nu}{\operatorname{argmax}}\, L(\nu; x_1, \ldots, x_N) = \underset{\nu}{\operatorname{argmax}} \sum_{n=1}^{N} \ell(\nu; x_n)$$

where $\ell(\nu; x_n) = \log f_X(x_n;\nu)$.

(a) One particular draw of data for $N = 50$ is variable `x` in the file `hw06p6a.mat`. Plot the log likelihood function, and report the MLE for ν. Your MLE will of course be approximate, but make sure yours is accurate to within 10^{-2} to the true MLE. I will give you a hint here and tell you that the true value of ν is somewhere in the interval $[0, 5]$.

(b) The file `hw06p6b.mat` contains a matrix `X`. This is an $N \times Q$ matrix, where $N = 50$ and $Q = 1000$; each entry is an independent Cauchy random variable with $\nu_0 = 3$. Treating each column of `X` as a single draw of the data for $N = 50$, compute the sample mean, sample median, and MLE for each column. From these, report the empirical mean squared error (by averaging $(\hat{\nu} - \nu_0)^2$ over all Q trials) for each of the three estimators.

(c) Find an integral expression for the expected log likelihood function $e(\nu) = \mathbb{E}[\ell(\nu; X)]$ when X has Cauchy density $f_X(x;\nu_0)$ as in (VI.1). Your expression should have the form

$$e(\nu) = \int_{-\infty}^{\infty} (\text{something that depends on } x, \nu, \nu_0)\ \underset{\cdot}{\mathrm{x}}.$$

Compute $e(\nu)$ for $\nu_0 = 3$ for 250 equally spaced values of ν between 0 and 5. You can do this using numerical integration (the `integral` function in MATLAB or `scipy.integrate.quad` in Python). Make a plot of $e(\nu) = \mathbb{E}[\ell(\nu; X)]$.

(d) Plot, overlayed on the same axes, the (renormalized) log likelihood functions $\frac{1}{N}\sum_{n=1}^{N} \ell(\nu; x_n)$ as a function of $\nu \in [0, 5]$ for each of the first 10 columns of `X` from part (b). On top of this, plot $e(\nu) = \mathbb{E}[\ell(\nu; X)]$ from part (c) as a dotted line.

Solution

Question (a)

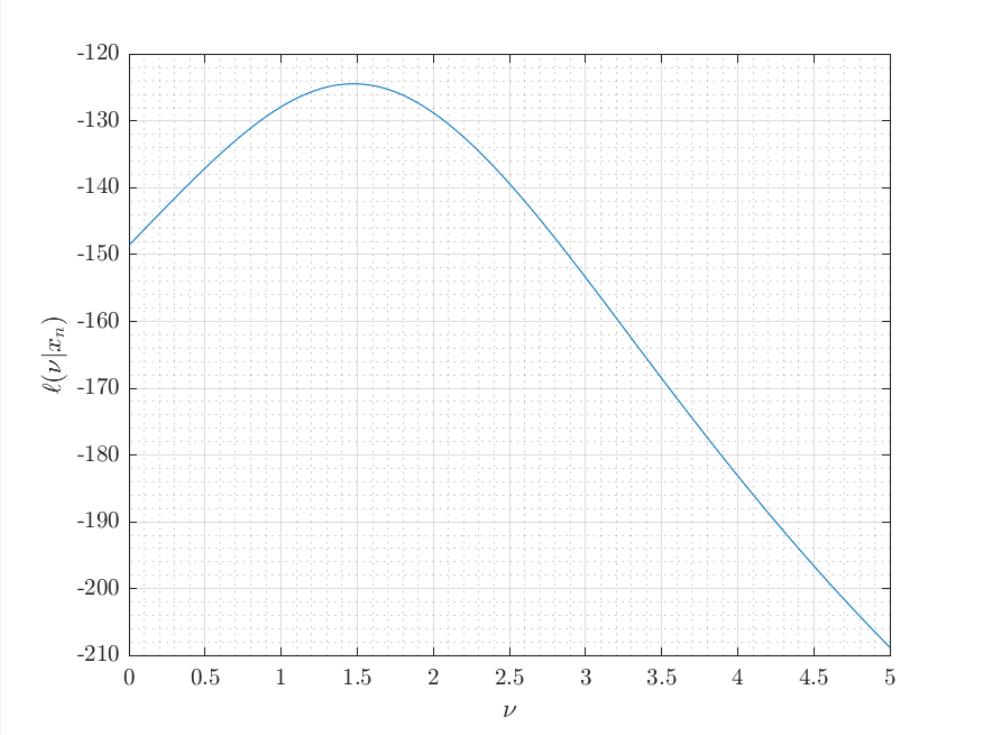
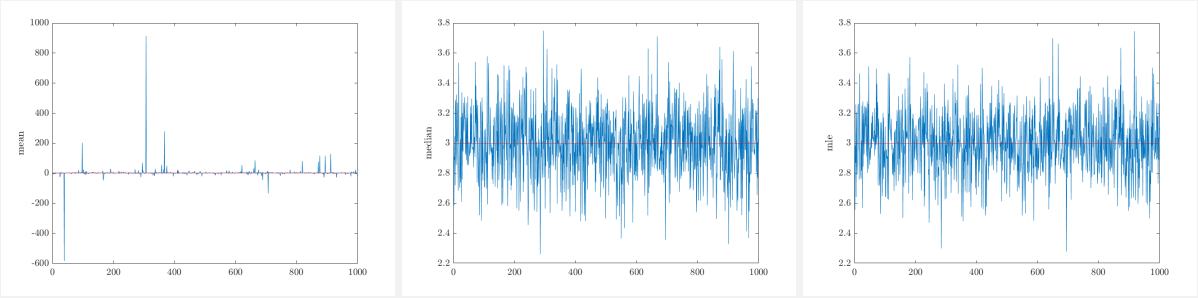


Figure 3: Log likelihood function plot for $\nu \in [0, 5]$.

The concave curve is maximum at the value of $\nu = 1.4743$.

Question (b)



(a) Mean (b) Median (c) MLE

Figure 4: Mean, median, and MLE for each columns of the data array.

The empirical mean squared error were

- $\hat{\nu}_{mn}$: 1411.2
- $\hat{\nu}_{md}$: 0.0501
- $\hat{\nu}_{mle}$: 0.0404

Question (c)

$$\mathbb{E}\left[\ell(\theta|X)\right] = \int_{-\infty}^{\infty} \frac{\ln\left\{\pi[1+(x-\nu)^2]\right\}}{\pi[1+(x-\nu_0)^2]} dx \tag{VI.2}$$

The plot is as follows

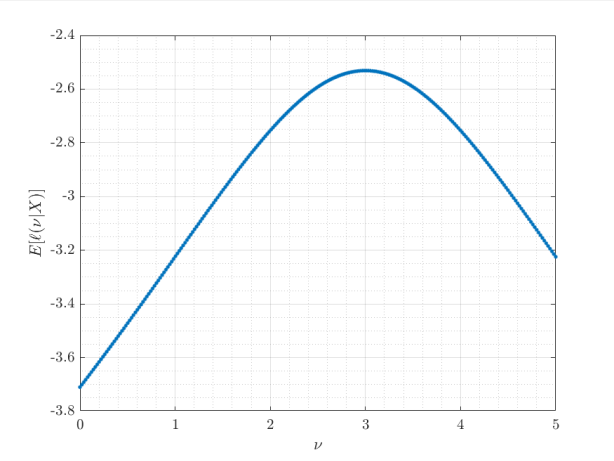


Figure 5: The expected log likelihood function plot.

If $X = X_1, \ldots, X_N$ where $N = 50$ we would have

$$\mathbb{E}\left[\ell(\theta|X)\right] = \int_{-\infty}^{\infty} N \frac{\ln\left\{\pi[1+(x-\nu)^2]\right\}}{\pi[1+(x-\nu_0)^2]} dx \tag{VI.3}$$

Question (d)

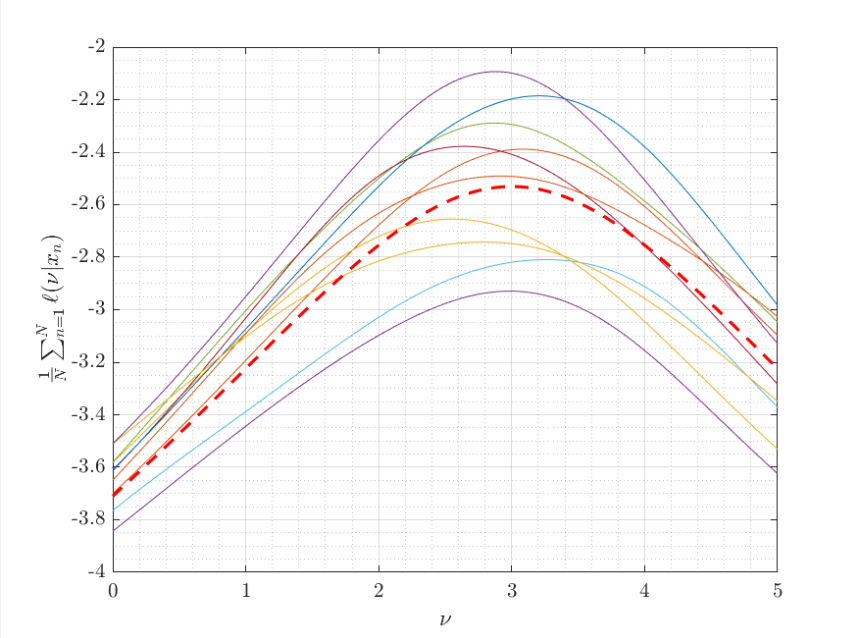


Figure 6: Overlay of expected log likelihood and normalized log likelihood functions.