Math Foundations of ML, Fall 2022

Homework #6

Due Monday November 14, at 5:00pm ET

As stated in the syllabus, unauthorized use of previous semester course materials is strictly prohibited in this course.

1. 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

$$F_Z(u|X = \beta) = \mathrm{P}\left(Z \leq u|X = \beta\right) = \mathrm{P}\left(Y \leq u - \beta|X = \beta\right).$$

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.
Since $Z = X + Y$, we have

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{\infty} F_Z(z|X = \beta) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} \mathrm{P}\left(Z \leq z|X = \beta\right) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} \mathrm{P}\left(Y \leq z - \beta|X = \beta\right) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} F_Y(z - \beta|X = \beta) f_X(\beta) d\beta. \end{aligned}$$

and thus

$$\begin{aligned} f_Z(z) = \frac{d}{dz} F_Z(z) &= \frac{d}{dz} \int_{-\infty}^{\infty} F_Y(z - \beta|X = \beta) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} \frac{d}{dz} F_Y(z - \beta|X = \beta) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} f_Y(z - \beta|X = \beta) f_X(\beta) d\beta \\ &= \int_{-\infty}^{\infty} f_{X,Y}(\beta, z - \beta) d\beta. \end{aligned}$$

If X and Y are independent,

$$f_Z(z) = \int_{-\infty}^{\infty} f_{X,Y}(\beta, z - \beta) d\beta = \int_{-\infty}^{\infty} f_X(\beta) f_Y(z - \beta) d\beta,$$

which is the convolution of the PDFs of X and Y.

2. 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.)

Solution.

We know that adding independent random variables is equivalent to convolving their PDFs. We realize $b_0(x) = f_X(x)$ is the 0th order B-spline, so $Y = X_1 + X_2 + X_3$ will have a PDF $f_Y(y) = (b_0 * b_0 * b_0)(y) = b_2(y)$, which is explicitly given below:

$$f_Y(y) = \begin{cases} (y+3/2)^2/2, & -3/2 \le y \le -1/2 \\ -y^2 + 3/4, & -1/2 \le y \le 1/2 \\ (y-3/2)^2/2, & 1/2 \le y \le 3/2 \\ 0, & |y| > 3/2 \end{cases}$$

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

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

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 $G \sim \text{Normal}(0, \sigma^2)$, 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.*)

Solution.

We first derive the MGF of X:

$$\begin{aligned} \varphi_X(t) &= \mathrm{E}\left[e^{tX}\right] \\ &= \int_{-1/2}^{1/2} e^{tx} dx \\ &= \frac{1}{t}\left(e^{\frac{t}{2}} - e^{-\frac{t}{2}}\right) \\ &= \frac{2}{t}\left(\frac{e^{\frac{t}{2}} - e^{-\frac{t}{2}}}{2}\right) \\ &= \frac{2}{t}\sinh\frac{t}{2}. \end{aligned}$$

Then we derive the MGF of Y_N:

$$\begin{aligned}
\varphi_{Y_N}(t) &= \mathrm{E}\left[e^{tY_N}\right] \\
&= \mathrm{E}\left[e^{\frac{t}{\sqrt{N}}\sum_{n=1}^{N} X_n}\right] \\
&= \mathrm{E}\left[\prod_{n=1}^{N} e^{\frac{t}{\sqrt{N}} X_n}\right] \\
&= \prod_{n=1}^{N} \mathrm{E}\left[e^{\frac{t}{\sqrt{N}} X_n}\right] \\
&= \mathrm{E}\left[e^{\frac{t}{\sqrt{N}} X}\right]^N \\
&= \left(\phi_X\left(\frac{t}{\sqrt{N}}\right)\right)^N \\
&= \left(\frac{2\sqrt{N}}{t}\sinh\frac{t}{2\sqrt{N}}\right)^N.
\end{aligned}$$

Please see “P2.ipynb” for the code and Figure 1 for the plot of each ϕ_{Y_N} and ϕ_G. From the plot we can conclude that $\varphi_{Y_N} \to \varphi_G$ as $N \to \infty$.

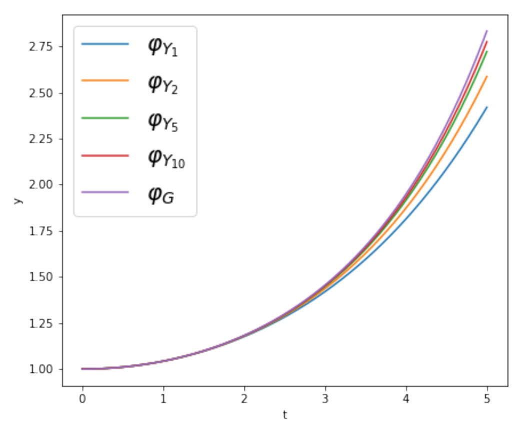

Figure 1: Plots of ϕ_{Y_N} and ϕ_G

Indeed, we can prove that $\varphi_{Y_N} \to \varphi_G$ as $N \to \infty$ rigorously. Since the Taylor expansion of sinh function is

$$\sinh(x) = x + \frac{x^3}{3!} + \frac{x^5}{5!} + \cdots = \sum_{n=0}^{\infty} \frac{x^{2n+1}}{(2n+1)!},$$

then we have

$$\frac{1}{x}\cdot\sinh(x) = 1 + \frac{x^2}{3!} + \frac{x^4}{5!} + \cdots = \sum_{n=0}^{\infty} \frac{x^{2n}}{(2n+1)!}.$$

Thus, we have

$$
\begin{aligned}
\lim_{N\to\infty} \varphi_{Y_N}(t) &= \lim_{N\to\infty} \left(\frac{2\sqrt{N}}{t} \sinh \frac{t}{2\sqrt{N}} \right)^N \\
&= \lim_{N\to\infty} \left(1 + \frac{\left(\frac{t}{2\sqrt{N}}\right)^2}{3!} \right)^N && \text{(Removed higher-order terms)} \\
&= \lim_{N\to\infty} \left(1 + \frac{t^2}{24N} \right)^N \\
&= \lim_{N\to\infty} \left(1 + \frac{1}{\frac{24N}{t^2}} \right)^{\frac{24N}{t^2}\cdot\frac{t^2}{24}} \\
&= e^{\frac{t^2}{24}} = \varphi_G(t).
\end{aligned}
$$

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

$$
\mathrm{P}\left(Z > u\right) = \mathrm{P}\left(\phi(Z) > \phi(u)\right).
$$

Use $\phi(z) = e^{tz}$ and the Markov inequality to derive a bound on $\mathrm{P}\left(Z_N > u\right)$, 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 Chebsyshev inequality.

Solution.

We first derive the MGF of Z_N:

$$
\begin{aligned}
\phi_{Z_N}(t) &= \mathrm{E}\left[e^{tZ_N}\right] \\
&= \mathrm{E}\left[e^{\frac{t}{N}\sum_{n=1}^{N} X_n}\right] \\
&= \mathrm{E}\left[\prod_{n=1}^{N} e^{\frac{t}{N}X_n}\right] \\
&= \prod_{n=1}^{N} \mathrm{E}\left[e^{\frac{t}{N}X_n}\right] \\
&= \mathrm{E}\left[e^{\frac{t}{N}X}\right]^N \\
&= \left(\phi_X\left(\frac{t}{N}\right)\right)^N \\
&= \left(\frac{2N}{t} \sinh \frac{t}{2N}\right)^N
\end{aligned}
$$

Then we derive the general Markov bound on Z_N:

$$\begin{aligned}
\mathrm{P}\left(Z_N > u\right) &\leq \frac{1}{e^{tu}}\,\mathrm{E}\left[e^{tZ_N}\right] \\
&= e^{-tu}\left(\frac{2N}{t}\sinh\frac{t}{2N}\right)^N
\end{aligned}$$

Choosing $t = 4u/N$ yields:

$$\mathrm{P}\left(Z_N > u\right) \leq e^{-\frac{4u^2}{N}}\left(\frac{N^2}{2u}\sinh\frac{2u}{N^2}\right)^N$$

We now derive the Chebyshev bound on Z_N:

$$\begin{aligned}
\mathrm{P}\left(|Z_N| > u\right) &\leq \frac{\mathrm{var}\left[Z_N\right]}{u^2} \\
&= \frac{\mathrm{var}\left[\frac{1}{N}\sum_{n=1}^N X_n\right]}{u^2} \\
&= \frac{\sum_{n=1}^N \mathrm{var}\left[X_n\right]}{N^2u^2} \\
&= \frac{1}{12Nu^2}
\end{aligned}$$

Apply the fact that Z_N is symmetrically distributed across the origin to get our final, tighter bound:

$$\begin{aligned}
\mathrm{P}\left(Z_N > u\right) &= \frac{1}{2}\,\mathrm{P}\left(|Z_N| > u\right) \\
&\leq \frac{1}{24Nu^2}
\end{aligned}$$

Please see "P2.ipynb" for the code and Figure 2 for the comparisons of the two bounds. When N is small, we see that the Markov bound is tighter bound for all u small and large enough. However, as N increases, the Markov bound loosens while the Chebyshev bound is tighter for all u large enough.

3. Let $Z_1, \ldots, Z_N$ 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 $X = \boldsymbol{x}$, 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

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

)

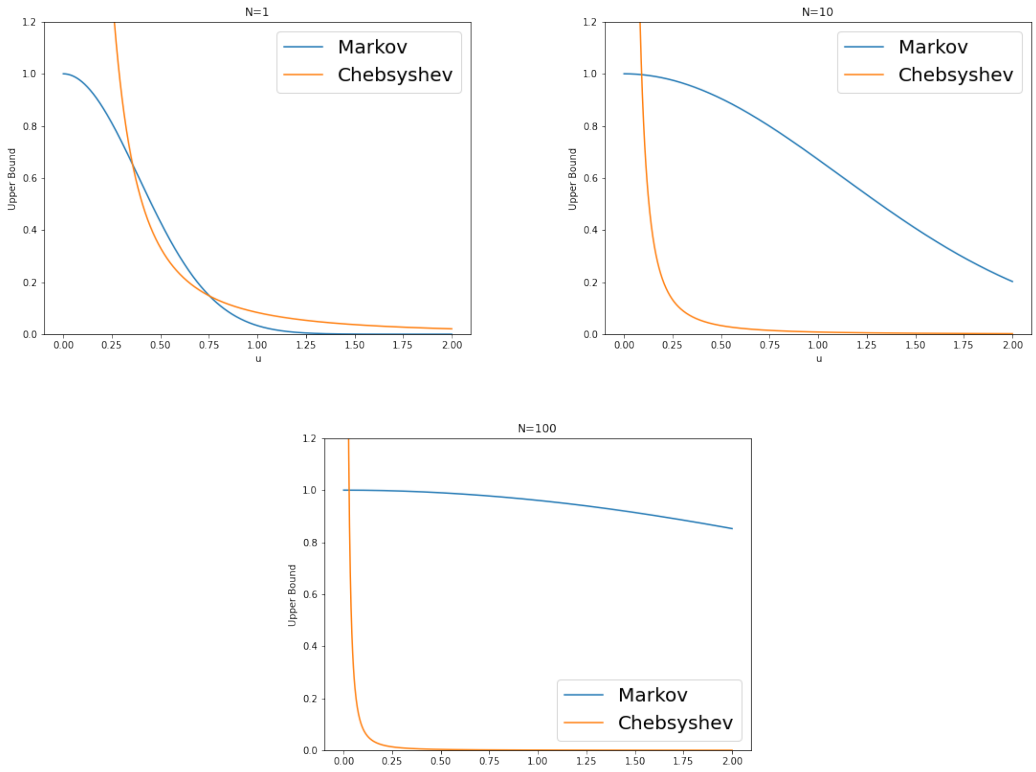

Figure 2: Comparisons of Markov and Chebyshev bounds for different N.

Solution.
Note that

$$f_{X_1}(x_1) \sim \mathrm{N}(0,1),$$

and

$$f_{X_k}(x_k|x_{k-1}) \sim \mathrm{N}(ax_{k-1},1), \quad \text{for all } k>1.$$

Thus,

$$\begin{aligned}\widehat{a}_{mle} &= \arg\max_{a\in\mathbb{R}} \ell(a;x_1,\ldots,x_N)\\ &= \arg\max_{a\in\mathbb{R}} -\frac{N}{2}\log(2\pi) - \frac{1}{2}x_1^2 - \frac{1}{2}\sum_{k=2}^{N}(x_k - ax_{k-1})^2\end{aligned}$$

Taking the 1st derivative of ℓ with respect to a and setting it equal to 0, we get

$$\widehat{a}_{mle} = \frac{\sum_{k=2}^{N} x_k x_{k-1}}{\sum_{k=2}^{N} x_{k-1}^2}.$$

Now taking the 2nd derivative of ℓ with respect to a, we have

$$\ell''(a;x_1,\ldots,x_N) = -\sum_{k=2}^{N} x_{k-1}^2 \leq 0.$$

Thus, we can conclude that $\widehat{a}_{mle} = \arg\max_{a\in\mathbb{R}} \ell(a; x_1, \ldots, x_N)$.

4. 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 $\boldsymbol{A}$ be a $M \times N$ matrix. We have

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

We will make observation of the random vector

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

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

Solution.
Since

$$R_{xy} = \mathrm{E}[XY^{\mathrm{T}}] = \mathrm{E}[X(\boldsymbol{A}X + E)^{\mathrm{T}}] = \mathrm{E}[XX^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}] = R_x\boldsymbol{A}^{\mathrm{T}},$$

and

$$\begin{aligned} R_y &= \mathrm{E}[YY^{\mathrm{T}}] = \mathrm{E}[(\boldsymbol{A}X + E)(\boldsymbol{A}X + E)^{\mathrm{T}}] \\ &= \boldsymbol{A}\,\mathrm{E}[XX^{\mathrm{T}}]\boldsymbol{A}^{\mathrm{T}} + \mathrm{E}[EE^{\mathrm{T}}] = \boldsymbol{A}R_x\boldsymbol{A}^{\mathrm{T}} + R_e, \end{aligned}$$

then we have

$$cov\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} \mathrm{E}[XX^{\mathrm{T}}] & \mathrm{E}[XY^{\mathrm{T}}] \\ \mathrm{E}[YX^{\mathrm{T}}] & \mathrm{E}[YY^{\mathrm{T}}] \end{bmatrix} = \begin{bmatrix} R_x & R_x\boldsymbol{A}^{\mathrm{T}} \\ \boldsymbol{A}R_x & \boldsymbol{A}R_x\boldsymbol{A}^{\mathrm{T}} + R_e \end{bmatrix}.$$

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

Solution.
In this problem, X is hidden, and Y is observed. We can write the MMSE of X given $Y = y$ as

$$\begin{aligned} \hat{\boldsymbol{x}}_{MMSE} &= R_{yx}^{\mathrm{T}} R_y^{-1} y \\ &= (\boldsymbol{A}R_x)^{\mathrm{T}}(\boldsymbol{A}R_x\boldsymbol{A}^{\mathrm{T}} + R_e)^{-1}y \\ &= R_x\boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A}R_x\boldsymbol{A}^{\mathrm{T}} + R_e)^{-1}y. \end{aligned}$$

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

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

Solution.

$$\begin{aligned}
\hat{\boldsymbol{x}}_{MMSE} &= R_x \boldsymbol{A}^{\mathrm{T}}(\boldsymbol{A} R_x \boldsymbol{A}^{\mathrm{T}} + R_e)^{-1}\boldsymbol{y} \\
&= \sigma_x^2 \boldsymbol{A}^{\mathrm{T}}(\sigma_x^2 \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \sigma_e^2 \mathbf{I})^{-1}\boldsymbol{y} \\
&= \sigma_x^2 \boldsymbol{V}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}}(\sigma_x^2 \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}} + \sigma_e^2 \mathbf{I})^{-1}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{U}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)\boldsymbol{U}^{\mathrm{T}}\right)^{-1}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^{\top}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} \\
&= \boldsymbol{V}\boldsymbol{\Sigma}^{\top}\left(\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{\top} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{y} \\
&= \sum_{n=1}^{R}\frac{\sigma_n}{\sigma_n^2 + \frac{\sigma_e^2}{\sigma_x^2}}\langle \boldsymbol{U}_n, \boldsymbol{y}\rangle \boldsymbol{v}_n
\end{aligned}$$

where σ_n denotes the n_{th} largest singular value of $\boldsymbol{A}$ and $\boldsymbol{U}_n$ the corresponding left singular vector.
Therefore, $\alpha_n = \frac{\sigma_n}{\sigma_n^2 + \frac{\sigma_e^2}{\sigma_x^2}}\langle \boldsymbol{U}_n, \boldsymbol{y}\rangle$ for $1 \le n \le R$, and $\alpha_n = 0$ for all $R < n \le N$.

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

$$\begin{aligned}
\mathrm{E}[\|\hat{\boldsymbol{x}}_{MMSE} - X\|_2^2] &= trace\left(\boldsymbol{R}_x - \boldsymbol{R}_{yx}^{\mathrm{T}}\boldsymbol{R}_y^{-1}\boldsymbol{R}_{yx}\right) \\
&= \sigma_x^2 \operatorname{trace}\left(\mathbf{I} - \boldsymbol{A}^{\mathrm{T}}\left(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{A}\right) \\
&= \sigma_x^2 \operatorname{trace}\left(\mathbf{I} - \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}}\left(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}^{\mathrm{T}} + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\right) \\
&= \sigma_x^2 \operatorname{trace}\left(\mathbf{I} - \boldsymbol{V}\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}^2 + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\right) \\
&= \sigma_x^2 \operatorname{trace}(\mathbf{I}) - \sigma_x^2 \operatorname{trace}\left(\left(\boldsymbol{\Sigma}^2 + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{\Sigma}\right) \\
&= N\sigma_x^2 - \sigma_x^2 \operatorname{trace}\left(\left(\boldsymbol{\Sigma}^2 + \frac{\sigma_e^2}{\sigma_x^2}\mathbf{I}\right)^{-1}\boldsymbol{\Sigma}^2\right) \\
&= N\sigma_x^2 - \sigma_x^2 \sum_{n=1}^{N}\frac{\sigma_n^2}{\sigma_n^2 + \frac{\sigma_e^2}{\sigma_x^2}}
\end{aligned}$$

where we make use of the identities $\operatorname{trace}(\boldsymbol{P} + \boldsymbol{Q}) = \operatorname{trace}(\boldsymbol{P}) + \operatorname{trace}(\boldsymbol{Q})$ and $\operatorname{trace}(\boldsymbol{P}\boldsymbol{Q}) = \operatorname{trace}(\boldsymbol{Q}\boldsymbol{P})$ if both $\boldsymbol{P}\boldsymbol{Q}$ and $\boldsymbol{Q}\boldsymbol{P}$ exist.

5. Let $\boldsymbol{A}$ 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 $\boldsymbol{R}_e$. Suppose we observe

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

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

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

Solution.

Y is a Gaussian random vector in $\mathbb{R}^M$:

$$Y \sim \mathrm{N}(\boldsymbol{A}\boldsymbol{\theta}_0, \boldsymbol{R}_e).$$

The mean of Y depends 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.)

Solution.

The maximum likelihood estimate of $\boldsymbol{\theta}_0$ can be found as follows:

$$\begin{aligned}
\widehat{\boldsymbol{\theta}}_0 &= \arg\max_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} L(\boldsymbol{\theta}_0; y) \\
&= \arg\max_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} \ell(\boldsymbol{\theta}_0; y) \\
&= \arg\max_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} \log((2\pi)^{-M/2}(\det \boldsymbol{R}_e)^{-1/2} \exp(-(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)^T \boldsymbol{R}_e^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)/2)) \\
&= \arg\max_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} -\frac{M}{2}\log(2\pi) + \frac{1}{2} log(\det \boldsymbol{R}_e^{-1}) - \frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)^T \boldsymbol{R}_e^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0) \\
&= \arg\max_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)^T \boldsymbol{R}_e^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0) \\
&= \arg\min_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} \|\boldsymbol{R}_e^{-1/2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)\|_2^2.
\end{aligned}$$

This can be solved as a least-squares problem

$$\arg\min_{\boldsymbol{\theta}_0 \in \mathbb{R}^N} \|(\boldsymbol{b} - \boldsymbol{H}\boldsymbol{\theta}_0)\|_2^2$$

with $\boldsymbol{b} = \boldsymbol{R}_e^{-1/2}\boldsymbol{y}$ and $\boldsymbol{H} = \boldsymbol{R}_e^{-1/2}\boldsymbol{A}$.

Thus

$$\widehat{\boldsymbol{\theta}}_0 = (\boldsymbol{H}^T\boldsymbol{H})^{-1}\boldsymbol{H}^T\boldsymbol{b} = (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{y}.$$

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

Solution.

$\hat{\boldsymbol{\Theta}}$ is a Gaussian random vector in $\mathbb{R}^N$ with mean

$$\begin{aligned}
\mathrm{E}[\hat{\boldsymbol{\Theta}}] &= \mathrm{E}[(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}Y] \\
&= \mathrm{E}[(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}(\boldsymbol{A}\boldsymbol{\theta}_0 + E)] \\
&= \mathrm{E}[(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A}\boldsymbol{\theta}_0 + (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}E] \\
&= \mathrm{E}[(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})\boldsymbol{\theta}_0 + (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}E] \\
&= \mathrm{E}[\boldsymbol{\theta}_0] + (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}^{-1}\,\mathrm{E}[E] \\
&= \boldsymbol{\theta}_0.
\end{aligned}$$

Let $\hat{\mathbf{\Theta}} = \boldsymbol{S}Y$ where $\boldsymbol{S} = (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}$, then we have

$$
\begin{aligned}
Var[\hat{\mathbf{\Theta}}] &= Var[\boldsymbol{S}Y] \\
&= \boldsymbol{S}Var[Y]\boldsymbol{S}^T \\
&= \boldsymbol{S}\boldsymbol{R}_e\boldsymbol{S}^T \\
&= (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{R}_e((\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{R}_e^{-1})^T \\
&= (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}\boldsymbol{A}^T(\boldsymbol{R}_e^{-1})^T\boldsymbol{A}((\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1})^T \\
&= (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})((\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^T)^{-1} \\
&= (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})(\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1} \\
&= (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}.
\end{aligned}
$$

Thus, we have

$$\hat{\mathbf{\Theta}} \sim \mathrm{N}(\boldsymbol{\theta}_0, (\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}).$$

$\hat{\mathbf{\Theta}}$ is unbiased since $\mathrm{E}[\hat{\mathbf{\Theta}}] = \boldsymbol{\theta}_0$.

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

Solution.

$$
\begin{aligned}
MSE(\hat{\mathbf{\Theta}}) &= \mathrm{E}[\|\hat{\mathbf{\Theta}} - \boldsymbol{\theta}_0\|_2^2] \\
&= trace(\boldsymbol{R}) + \|\,\mathrm{E}[\hat{\mathbf{\Theta}}] - \boldsymbol{\theta}_0\|_2^2 \\
&= trace((\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}).
\end{aligned}
$$

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

Solution.

The Fisher information matrix $\boldsymbol{J}(\boldsymbol{\theta}_0)$ is computed as below:

$$
\begin{aligned}
\boldsymbol{s}(\boldsymbol{\theta}_0; \boldsymbol{y}) &= \nabla_{\boldsymbol{\theta}_0}\ell(\boldsymbol{\theta}_0; \boldsymbol{y}) \\
&= \nabla_{\boldsymbol{\theta}_0}(-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)^T\boldsymbol{R}_e^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)) \\
&= \boldsymbol{A}^T\boldsymbol{R}_e^{-1}(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0),
\end{aligned}
$$

$$
\begin{aligned}
\boldsymbol{J}(\boldsymbol{\theta}_0) &= \mathrm{E}[\boldsymbol{s}(\boldsymbol{\theta}_0; \boldsymbol{y})\boldsymbol{s}(\boldsymbol{\theta}_0; \boldsymbol{y})^T] \\
&= \boldsymbol{A}^T\boldsymbol{R}_e^{-1}\,\mathrm{E}[(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\theta}_0)^T]\boldsymbol{R}_e^{-1}\boldsymbol{A} \\
&= \boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A}.
\end{aligned}
$$

Since

$$trace(\boldsymbol{J}(\boldsymbol{\theta}_0)^{-1}) = trace((\boldsymbol{A}^T\boldsymbol{R}_e^{-1}\boldsymbol{A})^{-1}) = MSE(\hat{\mathbf{\Theta}}_{mle}),$$

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.

The Cramer-Rao lower bound is the minimum mean squared error any unbiased estimator can achieve. Here, MLE is the best unbiased estimator of $\boldsymbol{\theta}_0$ since it meets the lower bound.

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

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

Despite its simple definition, this is a strange animal. First of all, its mean is not defined, as the integral $\int x/(1+x^2)\, dx$ 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 (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

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

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 $[0, 5]$.

Solution.

The MLE for ν is $\widehat{\nu}_{mle} = 1.4743$. Please see "P6.ipynb" for the code and Figure 3 for the plot of the log likelihood function.

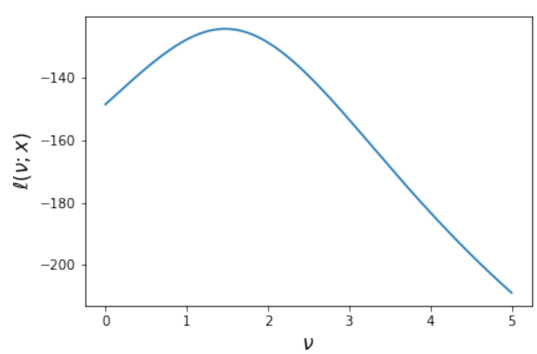

Figure 3: Plot of the log likelihood function.

(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.

Solution.
$\mathrm{MSE}(\widehat{\nu}_{mn}) = 1411.1503$, $\mathrm{MSE}(\widehat{\nu}_{md}) = 0.0501$ and $\mathrm{MSE}(\widehat{\nu}_{mle}) = 0.0404$. Please see "P6.ipynb" for the code.

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

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

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 $e(\nu) = \mathrm{E}[\ell(\nu; X)]$.

Solution.

$$e(\nu) = \int_{-\infty}^{\infty} \ell(\nu; x)\, f_X(x; \nu_0)\, dx = \int_{-\infty}^{\infty} \log\left(f_X(\nu; x)\right) f_X(x; \nu_0)\, dx.$$

Please see "P6.ipynb" for the code and Figure 4 for the plot of the expected log likelihood function.

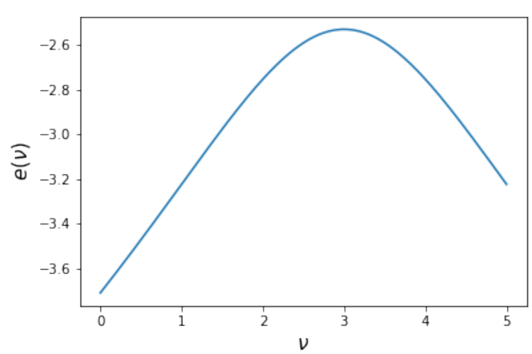

Figure 4: Plot of the expected log likelihood function.

(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 $e(\nu) = \mathrm{E}[\ell(\nu; X)]$ from part (c) as a dotted line.

Solution.
Please see "P6.ipynb" for the code and Figure 5 for the plot of the (renormalized) log likelihood functions.

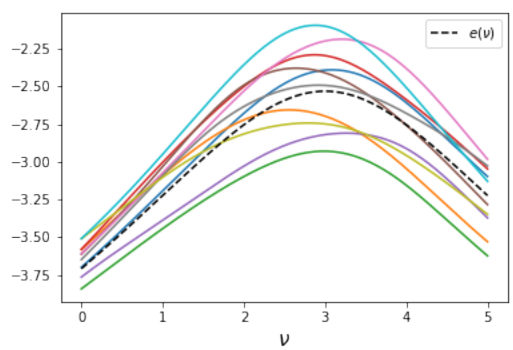

Figure 5: Plot of the (renormalized) log likelihood functions.