EE5137 Stochastic Processes: Problem Set 12
Assigned: 09/04/21, Due: Never

This problem set is not due but is examinable.

1. Suppose (X, Y) is a pair of random variables. Their joint density, depicted below, is constant in the shaded area and zero elsewhere.

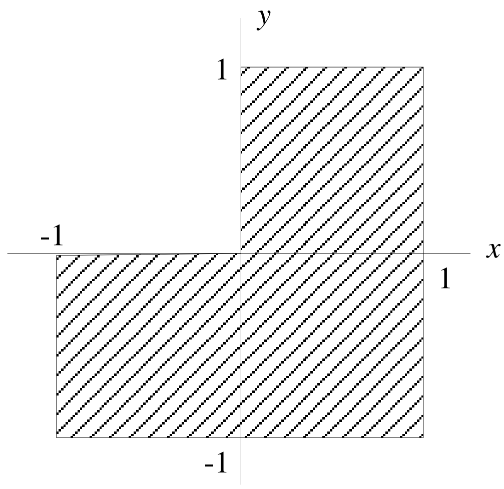

Figure 1: Joint density of (X, Y) for Question 1

(a) Determine $\hat{x}_{\mathrm{BLS}}(Y)$, the Bayes least squares estimate of X given Y.

Solution: We know that $\hat{x}_{\mathrm{BLS}}(Y) = \mathbb{E}[X|Y]$. We determine the conditional density by cutting the joint density at a particular value of y. If $0 < y \leq 1$, we get a uniform density between 0 and 1, which has mean 1/2. If $-1 \leq y \leq 0$, we get a uniform density between -1 and 1, which has mean 0. That is,

$$\hat{x}_{\mathrm{BLS}}(y) = \begin{cases} 1/2 & y > 0 \\ 0 & y \leq 0 \end{cases}.$$

(b) Determine λ_{BLS}, the error variance associated with your estimator in (a).

Solution: By iterated expectation, $\lambda_{\mathrm{BLS}} = \mathbb{E}[\lambda_{X|Y}]$. Now, $\lambda_{X|Y}$ is a discrete random variable that takes on two possible values, namely the variances for the two conditional densities we found in part (a). Using the fact that a random variable uniformly distributed between a and b has variance $(b-a)^2/12$, we see that these two values are 1/12 and 1/3, with probabilities $\Pr[Y > 0] = 1/3$ and $\Pr[Y \leq 0] = 2/3$, respectively. Thus

$$\lambda_{\mathrm{BLS}} = \mathbb{E}[\lambda_{X|Y}] = \frac{1}{12} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} = \frac{1}{4}.$$

(c) Consider the following "modified" cost function (corresponding to the cost of estimating x as $\hat{x}$):

$$C(x,\hat{x}) = \begin{cases} (x-\hat{x})^2 & x<0 \\ K(x-\hat{x})^2 & x \geq 0 \end{cases},$$

where $K > 1$ is a constant.

Determine $\hat{x}_{\text{MLS}}(Y)$, the associated Bayes estimate of X for this modified cost criterion. (MLS stands for *modified least squares.*)

Solution: For $y > 0$, we have $x \geq 0$, so the new cost function is a positive constant times the BLS cost function, and the optimal solution in the same. That is, $\hat{x}_{\text{MLS}}(y) = \hat{x}_{\text{BLS}}(y) = 1/2$ for $y > 0$.

In general,

$$\hat{x}_{\text{MLS}}(y) = \arg\min_a \underbrace{\int_{-\infty}^{\infty} C(x,a) f_{X|Y}(x|y)\,\mathrm{d}x}_{J(a)}.$$

For $y \leq 0$, $f_{X|Y}(x|y)$ is uniform on $[-1,1]$ We compute

$$J(a) = \int_{-1}^{0} (x-a)^2 \frac{1}{2}\,\mathrm{d}x + \int_0^1 K(x-a)^2 \frac{1}{2}\,\mathrm{d}x$$

Differentiating,

$$\frac{\mathrm{d}J}{\mathrm{d}a} = \frac{1-K}{2} + a(1+K).$$

It is clear that J has an absolute min at $a = \frac{1}{2}\frac{K-1}{K+1}$. We conclude that

$$\hat{x}_{\text{MLS}}(y) = \begin{cases} \dfrac{1}{2} & y > 0 \\ \dfrac{1}{2}\dfrac{K-1}{K+1} & y \leq 0 \end{cases}$$

(d) Give a brief intuitive explanation for why your answers to (a) and (c) are either the same or different.

Solution: When $y > 0$, the cost functions are scaled versions of each other, so our optimal estimators are the same. Now consider the case where $y \leq 0$: X is equally likely to be positive or negative, but there is a higher cost for our estimation error when the true value is non-negative. As K increases, the MLS estimator starts to "hedge its bets" and chooses larger values, tending towards $1/2$, at which point the possibility that X could be negative is almost completely ignored due to the (relatively) small cost of error that would be incurred.

2. You are given a coin and are allowed to toss it until you see the first head. You are then asked to estimate q, the probability of heads for this particular coin. Let Y be the number of times you see tails before the first head. We note that Y is distributed according to the geometric distribution

$$P_Y(y;q) = q(1-q)^y \quad y \in \mathbb{N} \cup \{0\}.$$

(a) Consider the maximum likelihood estimator of q. Is it unique? Is it efficient?

Solution: We optimize q over $[0,1]$. First consider the endpoints. We see that $q = 1$ achieves a maximum only when $y = 0$. $q = 0$ never achieves a maximum. Now say $0 < q < 1$ and $y > 0$. We take the first derivative, obtaining

$$\frac{\partial P_Y(y;q)}{\partial q} = (1-(y+1)q)(1-q)^{y-1}.$$

The derivative is zero when $q = 1/(y+1)$, positive when q is smaller than this quantity, and negative when q is larger. Thus the unique globally maximum is at

$$\hat{q}_{\mathrm{ML}}(y) = \frac{1}{y+1}.$$

This expression takes care of our special case at $y = 0$, so this is the ML estimator for all $y \geq 0$ By our argument, there is no other possibility for the ML estimator. It is unique.

Now, we look at the expected value of $\hat{q}_{\mathrm{ML}}(Y)$.

$$\mathbb{E}[\hat{q}_{\mathrm{ML}}(Y)] = \sum_{y=0}^{\infty} \frac{q(1-q)^y}{y+1} = q + \sum_{y=1}^{\infty} \frac{q(1-q)^y}{y+1} > q.$$

The ML estimate is biased, so it cannot be efficient.

(b) Your Bayesian friend proposes to treat the probability of heads as random variable X with a density that is uniform in the region $[0, 1]$. Find the Bayes least squares estimate for X. You can use the fact that

$$\int_0^1 x^k (1-x)^n \, \mathrm{d}x = \frac{n!k!}{(n+k+1)!}$$

Solution: We have

$$\begin{aligned}
\hat{x}_{\mathrm{BLS}}(Y) &= \mathbb{E}[X|Y] \\
&= \int x f_{X|Y}(x|y) \, \mathrm{d}x \\
&= \int x \frac{f_{Y|X}(y|x) f_X(x)}{f_Y(y)} \, \mathrm{d}x \\
&= \frac{\int x f_{Y|X}(y|x) f_X(x) \, \mathrm{d}x}{\int f_{Y|X}(y|x') f_X(x') \, \mathrm{d}x'} \\
&= \frac{\int_0^1 x f_{Y|X}(y|x) \, \mathrm{d}x}{\int_0^1 f_{Y|X}(y|x') \, \mathrm{d}x'} \\
&= \frac{\int_0^1 x^2 (1-x)^y \, \mathrm{d}x}{\int_0^1 x(1-x)^y \, \mathrm{d}x'} \\
&= \frac{\frac{y!2!}{(y+3)!}}{\frac{y!1!}{(y+2)!}} \\
&= \frac{2}{y+3}.
\end{aligned}$$

(c) We define relative bias to be the expected ratio of the true parameter to the estimated value, i.e.,

$$R_{\hat{x}} = \mathbb{E}\left[\frac{X}{\hat{x}(Y)}\right].$$

Just as what we've done with the usual additive bias, relative bias can be applied to non-random parameter estimation. An estimator with a relative bias of 1 is called *relatively unbiased*.

(i) Is the ML estimator you computed in part (a) relatively unbiased?

(ii) Is the BLS estimator you computed in part (a) relatively unbiased?

Solution: For the ML estimator,

$$r_{\hat{q}} = \mathbb{E}[q(Y+1)] = q\mathbb{E}[Y] + q = q\frac{1-q}{q} + q = 1.$$

For the BLS estimator,

$$\begin{aligned}
r_{\hat{x}} &= \mathbb{E}\left[X\frac{Y+3}{2}\right] \\
&= \frac{1}{2}(\mathbb{E}[XY] + 3\mathbb{E}[X]) \\
&= \frac{1}{2}(\mathbb{E}[\mathbb{E}[XY|X]] + 3\mathbb{E}[X]) \\
&= \frac{1}{2}(\mathbb{E}[X\mathbb{E}[Y|X]] + 3\mathbb{E}[X]) \\
&= \frac{1}{2}\left(\mathbb{E}\left[X\frac{1-X}{X}\right] + 3\mathbb{E}[X]\right) \\
&= \frac{1}{2} + \mathbb{E}[X] \\
&= 1.
\end{aligned}$$

Both are relatively unbiased.

3. (a) Let

$$f_Y(y;x) = \begin{cases} x & 0 \le y \le 1/x \\ 0 & \text{else} \end{cases}$$

for $x > 0$. Show that there exist no unbiased estimators $\hat{x}(Y)$ of x. (Note that because only $x > 0$ are possible values, an unbiased estimator need only be unbiased for $x > 0$ rather than all x.)

Solution: Assume towards a contradiction that $\hat{x}(Y)$ is unbiased, so that $\mathbb{E}[\hat{x}(Y)] = x$. That is, for all $x > 0$:

$$\int_0^{1/x} \hat{x}(y)x \, \mathrm{d}y = x.$$

Since $x \neq 0$, we can cancel x on both sides yielding

$$\int_0^{1/x} \hat{x}(y) \, \mathrm{d}y = 1.$$

Now, for any $b > a > 0$, we may write

$$\begin{aligned}
\int_a^b \hat{x}(y) \, \mathrm{d}y &= \int_0^b \hat{x}(y) \, \mathrm{d}y - \int_0^a \hat{x}(y) \, \mathrm{d}y \\
&= \int_0^{\frac{1}{1/b}} \hat{x}(y) \, \mathrm{d}y - \int_0^{\frac{1}{1/a}} \hat{x}(y) \, \mathrm{d}y \\
&= 1 - 1 = 0.
\end{aligned}$$

Since a and b are arbitrary positive numbers, $\hat{x}(y) = 0$ for $y > 0$. But now $x = \mathbb{E}[\hat{x}(Y)] = 0$, a contradiction.

(b) Let

$$f_Y(y;x) = \begin{cases} 1/x & 0 \le y \le x \\ 0 & \text{else} \end{cases}$$

for $x > 0$. Show that there is a unique unbiased estimator.

Solution: Let's say $\hat{x}(y) = g(y)$ is unbiased, and see what constraints this places on g. We have

$$x = \mathbb{E}[\hat{x}(Y)] = \int_0^x \frac{g(y)}{x}\, \mathrm{d}y.$$

Multiplying both sides by x, we get

$$\int_0^x g(y)\, \mathrm{d}y = x^2.$$

Next, we differentiate with respect to x, and obtain $g(x) = 2x$, for all $x > 0$. That's a pretty strict requirement for our estimator! The only thing that's not nailed down is $g(x)$ for $x \leq 0$, and that's not going to impact the variance at all, because $y > 0$ with probability 1. Thus $\hat{x}(y) = 2y$ is the unique unbiased estimator.

4. Suppose that for $i = 1, 2$,

$$Y_i = x + W_i$$

where x is an unknown but non-zero constant, and where W_1 and W_2 are statistically independent, zero-mean Gaussian random variables with

$$\begin{aligned} \mathrm{Var}(W_1) &= 1 \\ \mathrm{Var}(W_2) &= \begin{cases} 1 & x > 0 \\ 2 & x < 0 \end{cases}. \end{aligned}$$

(a) Calculate the Cramér-Rao bound for unbiased estimators of x based on an observation of $\mathbf{Y} = (Y_1, Y_2)$.

Solution: We first compute the Fisher information. In the calculation below, C_1 and C_2 are constants.

$$\begin{aligned} f_{\mathbf{Y}}(\mathbf{y}; x) &= \begin{cases} N(y_1; x, 1) N(y_2; x, 1) & x > 0 \\ N(y_1; x, 1) N(y_2; x, 2) & x < 0 \end{cases} \\ \log f_{\mathbf{Y}}(\mathbf{y}; x) &= \begin{cases} C_1 - \frac{1}{2}(y_1 - x)^2 - \frac{1}{2}(y_2 - x)^2 & x > 0 \\ C_2 - \frac{1}{2}(y_1 - x)^2 - \frac{1}{4}(y_2 - x)^2 & x < 0 \end{cases} \\ \frac{\partial \log f_{\mathbf{Y}}(\mathbf{y}; x)}{\partial x} &= \begin{cases} y_1 - x + y_2 - x & x > 0 \\ y_1 - x + \frac{1}{2}(y_2 - x) & x < 0 \end{cases} \\ \frac{\partial^2 \log f_{\mathbf{Y}}(\mathbf{y}; x)}{\partial x^2} &= \begin{cases} -2 & x > 0 \\ -\frac{3}{2} & x < 0 \end{cases} \\ J_{\mathbf{Y}}(x) &= -\mathbb{E}\left[\frac{\partial^2 \log f_{\mathbf{Y}}(\mathbf{Y}; x)}{\partial x^2}\right] = \begin{cases} 2 & x > 0 \\ \frac{3}{2} & x < 0 \end{cases} \end{aligned}$$

Now, the Cramér-Rao bound is

$$\lambda_{\hat{x}}(x) \geq \begin{cases} \frac{1}{2} & x > 0 \\ \frac{2}{3} & x < 0 \end{cases}.$$

(b) Show that a minimum variance unbiased estimator $\hat{x}_{\mathrm{MVU}}(Y)$ does not exist.

Hint: Consider the estimators

$$\hat{X}_1 = \frac{Y_1}{2} + \frac{Y_2}{2} \quad \textit{and} \quad \hat{X}_2 = \frac{2Y_1}{3} + \frac{Y_2}{3}.$$

Solution: Let's see how good the estimators in the hint are. First, we'll compute their expectations:

$$\mathbb{E}[\hat{x}_1(\mathbf{Y})] = \frac{1}{2}\mathbb{E}Y_1 + \frac{1}{2}\mathbb{E}Y_2 = \frac{1}{2}x + \frac{1}{2}x = x$$
$$\mathbb{E}[\hat{x}_2(\mathbf{Y})] = \frac{2}{3}\mathbb{E}Y_1 + \frac{1}{3}\mathbb{E}Y_2 = \frac{2}{3}x + \frac{1}{3}x = x.$$

So, these estimators are unbiased. Now let's look at their error variances:

$$\mathrm{Var}[\hat{x}_1(\mathbf{Y})] = \frac{1}{4}\mathrm{Var}(Y_1) + \frac{1}{4}\mathrm{Var}(Y_2) = \begin{cases} \frac{1}{2} & x > 0 \\ \frac{3}{4} & x < 0 \end{cases}$$
$$\mathrm{Var}[\hat{x}_2(\mathbf{Y})] = \frac{4}{9}\mathrm{Var}(Y_1) + \frac{1}{9}\mathrm{Var}(Y_2) = \begin{cases} \frac{5}{9} & x > 0 \\ \frac{2}{3} & x < 0 \end{cases}$$

Aha. It appears that $\hat{x}_1(\mathbf{Y})$ achieves the CRLB if $x > 0$, and $\hat{x}_2(\mathbf{Y})$ achieves the CRLB if $x < 0$. But we don't know the sign of x, and we can't get a single estimator that achieves the CRLB for all x from these two estimators.

---

5. (Optional) Let $\mathbf{Y} = [Y_1, Y_2, \ldots, Y_n]^\top$ be an n-dimensional random vector composed of independent scalar Gaussian random variables Y_i, each with unknown, non-random mean x_i and unit variance. Our goal is to construct an estimator $\hat{x}(\mathbf{Y})$ for the vector of the mean parameters $\mathbf{x} = [x_1, x_2, \ldots, x_n]^\top$. Here, we generalize the mean-square error cost function to vector parameters in the following way:

$$\mathrm{MSE}_{\hat{\mathbf{x}}}(\mathbf{x}) = \mathbb{E}\left[\|\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x}\|^2\right] = \mathbb{E}\left[(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x})^\top(\hat{\mathbf{x}}(\mathbf{Y}) - \mathbf{x})\right].$$

(a) Determine the maximum likelihood estimator $\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y})$.

Solution: Because the Gaussian random variables are independent, the joint density of $\mathbf{Y}$ factors into a product of Gaussian densities for each Y_i. Therefore, the ML estimator for the entire vector is composed of ML estimators for each component. The ML estimate of the i-th component of $\mathbf{x}$, i.e., the mean of the Gaussian random variable Y_i, is simply the observation Y_i. Therefore $\hat{\mathbf{x}}_{\mathrm{ML}}(\mathbf{y}) = \mathbf{y}$.

(b) Find the bias of the maximum likelihood estimator.

Solution: The ML estimator for the mean of a Gaussian random variable is unbiased. Hence, bias $\mathbf{b}_{\hat{\mathbf{x}}_{\mathrm{ML}}}(\mathbf{x}) = \mathbf{0}$.

(c) Determine $\mathrm{MSE}_{\hat{\mathbf{x}}_{\mathrm{ML}}}(\mathbf{x})$, the mean-square error (MSE) of the maximum likelihood estimator.

Solution: Given than the estimator is unbiased, the MSE is simply the variance. The variance is the sum of the variances along each component, i.e., the sum of the variances of Y_i. Therefore, $\mathrm{MSE}_{\hat{\mathbf{x}}_{\mathrm{ML}}}(\mathbf{x}) = n$.

(d) Reading through a statistics book, you find a highly curious estimator

$$\hat{x}_{\mathrm{JS}}(\mathbf{y}) = \mathbf{y} - (n-2)\frac{\mathbf{y}}{\|\mathbf{y}\|_2^2}.$$

Show that

$$\mathrm{MSE}_{\hat{x}_{\mathrm{JS}}}(\mathbf{x}) = \alpha \mathrm{MSE}_{\hat{\mathbf{x}}_{\mathrm{ML}}}(\mathbf{x}) + \beta\mathbb{E}\left[\frac{1}{\|\mathbf{Y}\|_2^2}\right],$$

and find α and β.

Hint: Use this special case of Stein's lemma:

$$\mathbb{E}\left[(\mathbf{x}-\mathbf{Y})^\top \frac{\mathbf{Y}}{\|\mathbf{Y}\|_2^2}\right] = -(n-2)\mathbb{E}\left[\frac{1}{\|\mathbf{Y}\|_2^2}\right].$$

In fact, the curious estimator $\mathrm{MSE}_{\hat{x}_{\mathrm{JS}}}(\mathbf{x})$ *is known as the James-Stein estimator. If you have time, read up on it on Wikipedia.*

Solution: This estimator is known as the James-Stein estimator. Its discovery in the 50's took the statistics community by surprise. Recall the linear algebra identity

$$\|\mathbf{u}+\mathbf{v}\|^2 = \|\mathbf{u}\|^2 + 2\langle \mathbf{u}, \mathbf{v}\rangle + \|\mathbf{v}\|^2$$

and let $\mathbf{u} = \mathbf{x} - \mathbf{y}$ and $\mathbf{v} = (n-2)\frac{\mathbf{y}}{\|\mathbf{y}\|^2}$. The middle term can be simplified immediately using the special case of Stein's lemma

$$\mathbb{E}\left[(\mathbf{x}-\mathbf{Y})^\top \frac{\mathbf{Y}}{\|\mathbf{Y}\|_2^2}\right] = -(n-2)\mathbb{E}\left[\frac{1}{\|\mathbf{Y}\|_2^2}\right].$$

Note that $\|\mathbf{u}\|^2 = \mathrm{MSE}_{\hat{\mathbf{x}}_{\mathrm{ML}}}(\mathbf{x})$. Combining terms, we have $\alpha = 1$ and $\beta = -(n-2)^2$.

(e) We say that an estimator $\hat{\mathbf{x}}$ dominates another estimator $\hat{\mathbf{x}}'$ under MSE if $\mathrm{MSE}_{\hat{\mathbf{x}}}(\mathbf{x}) \leq \mathrm{MSE}_{\hat{\mathbf{x}}'}(\mathbf{x})$ for all $\mathbf{x}$ and the inequality is strict for some $\mathbf{x}$. An estimator is *admissible* if no other estimator dominates it; otherwise it is *inadmissible*.

Show that the maximum likelihood estimator $\mathrm{MSE}_{\hat{\mathbf{x}}_{\mathrm{ML}}}$ is inadmissible for $n > 2$.

Solution: For $n > 2$, $\beta < 0$ and the MSE of the curious estimator (the James-Stein estimator) dominates the MSE of the ML estimator. Therefore, the ML estimator is inadmissible.

We conclude with a few remarks. The ML estimator above is the MVU estimator and we showed in (e) that it is inadmissible.

Isn't this a contradiction? No. The MVU must be unbiased, and, by allowing bias, the James-Stein estimator lowers the variance enough to improve the MSE uniformly. Interpreting the James-Stein estimator is easy if one takes a Bayesian perspective.

Consider the Bayesian setting where you treat each mean parameter x_i as an independent draw from some known distribution. Which estimator, $\hat{\mathbf{x}}_{\mathrm{ML}}$ or $\hat{\mathbf{x}}_{\mathrm{JS}}$, will the Bayesian parameter estimate more closely resemble? Bayesian inference in this setting would result in n independent parameter estimates and no data-sharing because the problems are completely independent. Data for y_i is independent of x_j for $j \neq i$. Similarly, the ML estimator only uses y_i to estimate x_i. The James-Stein estimator apparently uses all of the data to estimate each parameter.

From a Bayesian perspective, what is the James-Stein estimator not assuming about the mean parameters x_i? Apparently, the James-Stein estimator is not assuming independence between the x_i's. In fact, you can recover the James-Stein estimator with the following Bayesian model: let the x_i be independent draws from a Gaussian distribution with mean 0 and variance σ^2, where σ^2 itself is a random variable with a improper (i.e. unnormalizable) uniform distribution on $[-1, \infty)$ Negative variance? Weird. In fact, by this correspondence, it's clear that the estimator can be uniformly improved by changing the prior to be uniform on $[0, \infty)$.

The James-Stein estimator, while a mystery/paradox to orthodox statisticians, has a sensible Bayesian interpretation: the means are independent but related to each other and this dependence is modeled by a very disperse distribution (a Gaussian prior with a uniform prior on variance is related to the Student t-distribution).