35V3A1: Computational Aspects in Econometrics
2023-2024
Pieter Kleer
P.S.KLEER@TILBURGUNIVERSITY.EDU

Matlab Assignment

Deadline: Wednesday, January 3, 2024, 23:59 (CET)

General instructions

- This is group assignment and all group members are expected to contribute to the same degree. Group members are the ONLY ones that contribute to the assignment.[1] Every goup member should be able to explain what was handed in upon request.
- You are allowed to use functionality from all official Matlab toolboxes (that are available in the TiU student installation).
- Apart from correctly solving the problems, your submission is also assessed on other criteria, such as efficiency, hard coding, DRY, single responsibility, coding style & documentation, KISS. See the introductory slides of the Matlab module. As a rule of thumb, the more you exploit the functionality as explained in the theory of the Matlab module, the higher your grade. In particular, avoid for loops whenever possible and do not repeat code unnecessarily (e.g., if you have used something useful in an earlier question already). Define auxiliary functions if needed.
- Always add titles, labels, and/or legends to your figures to identify the figure and its contents. Add the names of the team members at the top of the files.
- If you have questions about the assignment, ask them in class or send me an e-mail.

What to hand in?

- A short report **(PDF)** in which you explain your solutions and a .m-file with your code. For both there is a **mandatory template** on Canvas (replace `[group_number]` by your group number in both).
  - If you are asked to produce figures, these have to be included in the report as well (apart from being outputted by your Matlab code).
  - Give a description of every solution in the report to illustrate you understood what you implemented. **No explanation** = **no points.** You should also add some comments to the Matlab file, so that its structure is clear, but use the report for a more elaborate explanation.
  - Give a short summary in the report of who did what for this assignment.
- All requested documents have to be handed in before the deadline mentioned above. Late submissions automically get grade 1.0 (no discussion possible).

[1] Aids like ChatGPT are forbidden. Fraud suspicions in this or other regards are sent to the Examination Board.

1 Optimal stopping problem

We consider a stopping game (played by a gambler) given by a collection of n boxes, where box i is equipped with a nonnegative (continuous) probability distribution $\mathcal{F}_i$ for $i = 1,\dots,n$. We place a realization v_i of a random variable $X_i \sim \mathcal{F}_i$ in box i and then close it.

We present the boxes to the gambler who gets to open them one by one in the order $(1,\dots,n)$. Upon opening box i, the gambler has to decide whether she accepts or refuses the value v_i. This decision is irrevocable: Either she accepts the value and the game stops, or she refuses the value and moves on to the next box. This means she only gets to keep the value of one of the boxes, and, furthermore, once she refuses the value of a box, this decision cannot be turned around. That is, she cannot go back to this box later on, if future values turn out to be disappointing.

The gambler knows the distributions $\mathcal{F}_1,\dots,\mathcal{F}_n$ but (of course) not the realized values $v_1,\dots,v_n$. Her goal is to come up with a good strategy specifying at which box to stop, so that the expected value she obtains is optimal (or at least decent). Here the expectation is taken with respect to the random variables $X_1,\dots,X_n$, which are assumed to be independent of each other. More informally, if the gambler has chosen a strategy on when to stop, and plays the game using this strategy many times, then she obtains an average payoff from this strategy (which is the average over the values that were selected in the individual rounds of the game). Her goal is to maximize this average.

In this assignment, we will implement ”threshold-based” strategies. The idea is that the gambler sets a threshold number t_i for every box, and selects the first realized value v_i whose value exceeds the threshold t_i, i.e., the first i for which $v_i \geq t_i$. These thresholds will be computed based on the distributional information $\mathcal{F}_1,\dots,\mathcal{F}_n$ that the gambler has at the start of the game.

1.1 Recursively defined thresholds

The optimal thresholds for the gambler can be determined by means of a backward recursion. First of all, note that, conditioned on the fact that the gambler refused the first $n-1$ boxes, it is always optimal to accept the value in the last box (if she also refuses box n, she gets nothing). The expected value she obtains from the last box is $\mathbb{E}[X_n]$.

Secondly, conditioned on the fact that she refuses the first $n-2$ boxes, what is the best strategy to follow for box $n-1$? The gambler knows that if she would refuse the value v_{n-1}, then her expected value obtained from the last box is $T_{n-1} = \mathbb{E}[X_n]$. Therefore, the optimal strategy at box $n-1$ is to accept v_i if and only if $v_{n-1} \geq T_{n-1}$ (i.e., if the value is higher than what she expects to get if she refuses the box and moves on to box n). Her expected payoff over the last two boxes is then[2]

$$T_{n-2} = \mathbb{E}_{\mathcal{F}_{n-1}}[\max\{X_{n-1}, T_{n-1}\}]$$

Now, T_{n-2} is the expected payoff the gambler obtains from the last two boxes, which is now used as threshold for box $n-2$. One can continue this reasoning all the way back to the first box. In general, the threshold for box i is given by

$$T_i = \mathbb{E}_{\mathcal{F}_{i+1}}[\max\{X_{i+1}, T_{i+1}\}] \tag{1}$$

[2] This is elementary probability theory (think a bit about it if you don’t see it right away).

for $i = 1, \dots, n-1$.

a) **[5 points]** Write a function that takes as input continuous probability density functions (pdf) $\mathcal{F}_1, \dots, \mathcal{F}_n$, and outputs the thresholds T_i for all boxes $i = 1, \dots, n$, as well as the expected payoff (which is T_0) of the gambler using this threshold strategy.
The way in which you generate the pdfs in part b) is related to your implementation here, so think about part b) simultaneously. Your function here should work for all continuous pdfs, not only those given in part b), though.

b) **[4 points]** Instantiate pdfs of the following probability distributions for $n = 9m = 27$ (so that $m = 3$):

$$\mathcal{F}_i = \begin{cases} \text{Half-normal}(\mu, \sigma) & \text{if } i = 1,2,3,10,11,12,\dots,n-8,n-7,n-6 \\ \text{Gamma}(a,b) & \text{if } i = 4,5,6,13,14,15,\dots,n-5,n-4,n-3 \\ \text{Unif}(\ell,u) & \text{if } i = 7,8,9,16,17,18,\dots,n-2,n-1,n \end{cases},^{3}$$

with $(\mu, \sigma, a, b, \ell, u) = (1,5,1,10,1,10)$. The first distribution is the Half-normal distribution with parameters μ and σ, the second one the Gamma distribution with shape parameters a and b, and the third one the uniform distribution over the interval $[\ell, u]$. Execute your function of part a) using these pdfs as input. *Hint: Define the pdfs as anonymous functions and choose an appropriate data structure to store them in (that serves as the input for the function in part a)).*

We will next construct some empirical evidence to illustrate that the thresholds in (1) are indeed optimal. We will do this by testing the performance of various other threshold-based strategies. We take the thresholds from (1) and multiply them with a given value $\alpha \in [0,2]$, i.e., we define the threshold vector $T^\alpha = (T_1^\alpha, T_2^\alpha, \dots, T_n^\alpha)$ by

$$T_i^\alpha = \alpha \cdot T_i$$

with T_i as in (1), for $i = 1, \dots, n$.

c) **[5 points]** Write a function that takes as input an (arbitrary) threshold vector $t = (t_1, \dots, t_n)$ and vector of realized values $(v_1, \dots, v_n)$. It should output the index i^* of the box the gambler chooses[4] and its value v_{i^*}.

d) **[2 points]** Write a function that takes as input integers n and d, and the parameters $(\mu, \sigma, a, b, \ell, u)$. It should output d vectors of realizations $(v_1, \dots, v_n)$ where $v_i \sim X_i$ with X_i distributed according to $\mathcal{F}_i$ as in part b) for $i = 1, \dots, n$.

e) **[3 points]** Generate $d = 1.000.000$ (one million) vectors of realizations using your function of part d). For every $\alpha \in \{j/20 : j = 0,1,\dots,40\}$ apply your function in part c) to the threshold vector T^α for all d realization vectors. Compute for every α the average of the value of the selected box of these d runs. Give a plot with the values of α on the x-axis, and the computed average payoffs under the threshold vectors T^α on the y-axis. *Explain in the report what you see in the figure (that also needs to be in the report).*

[3] That is, $(H,H,H,G,G,G,U,U,U,H,H,H,G,G,G,U,U,U,...,H,H,H,G,G,G,U,U,U)$ is the pattern of the distributions of the boxes.

[4] Using the "threshold-based" strategy based on the t_i as explained in the introduction.

f) **[1 point]** Print the message *"The difference between the optimal expected payoff ([A]) of the gambler and the simulated average optimal payoff ([B]) is [C]"*, where you replace

- [A] by T_0 as computed in part a),
- [B] by your answer of part e) for $\alpha = 1$,
- [C] by the absolute difference between these numbers.

1.2 Cost for waiting

We consider a variation of the setting in the introduction where there is a box-dependent cost that has to be paid at the moment that we stop. The cost of stopping at box i is c_i for $i = 1, \ldots, n$. We will consider threshold-based strategies of the form $t_i^\beta = \beta_i \cdot T_i$ for $i = 1, \ldots, n$ with $T = (T_1, \ldots, T_n)$ as in (1), and $\beta = (\beta_1, \ldots, \beta_n)$ a vector of nonnegative numbers. The goal is to find, for a given vector of costs $c = (c_1, \ldots, c_n)$ and d vectors of realizations, a vector $\beta = (\beta_1, \ldots, \beta_n)$ that maximizes the average utility of the gambler, which in this case is given by

$$u(\beta, T, c, v^{(1)}, \ldots, v^{(d)}) = \frac{1}{d} \sum_{j=1}^{d} v_{i^*(j)}^{(j)} - c(i^*(j)). \tag{2}$$

Here $i^*(j)$ is the index of the chosen box in the realization vector $v^{(j)} = (v_1^{(j)}, v_2^{(j)}, \ldots, v_n^{(j)})$ when applying the treshold-based strategy $t^\beta = (\beta_1 T_1, \beta_2 T_2, \ldots, \beta_n T_n)$. That is, the utility computes for every realization vector the difference between the chosen value and the cost, and averages these differences over the d realization vectors. For given $T, c, v^{(1)}, \ldots, v^{(d)}$, we want to solve the maximization problem

$$\max_{\beta} u(\beta, T, c, v^{(1)}, \ldots, v^{(d)}) \tag{3}$$

i.e., we want to compute parameters $\beta_1, \ldots, \beta_n$, so that the utility of the gambler is maximized.

g) **[2 points]** Write a function that takes as input β, T, c and $v^{(1)}, \ldots, v^{(d)}$, and outputs the utility in (2).
Although we will later use the optimal thresholds as computed by the function in part a), in the current function T is just a generic threshold vector (same in part h)).

h) **[3 points]** Write a function that takes as input $T, c, v^{(1)}, \ldots, v^{(d)}$ and returns an optimal vector β to (3), together with its average utility.
Use FMINSEARCH() *in your function to do the minimization over β. This function might get stuck in a local minimum, but we ignore that in this assignment.*

We will next test the function from part h) on some input.

i) **[2 points]** Take the following input:

- ○ $n = 25$ and $\mathcal{F}_i = \text{Unif}(0, L)$ for every $i = 1, \ldots, n$ with $L = 15$ (i.e., we have I.I.D. distributions).

- ○ $d = 10.000$ (ten thousand)
- ○ $c : \{1,\dots,n\} \to \mathbb{R}$ defined by $c(i) = (i/n) \cdot (L/2)$.

Generate the optimal thresholds T_i for $i = 1,\dots,n$ using your function in part a) for the given uniform distributions. Use the vector $T = (T_1,\dots,T_n)$, together with c and d generated realization vectors, to run your function from part h).

1.3 Order statistics-based thresholds

Given n I.I.D. random variables distributed according to $\mathcal{F}$, the k-th order statistic is the distribution of the k-th smallest value of a statistical sample.[5] In particular, if $X_1,\dots,X_n$ are distributed according to $\mathcal{F}$, then $X_{(k)}$ has cumulative distribution function (cdf) given by

$$\mathbb{P}(X_{(k)} \leq x) = \sum_{j=k}^{n} \binom{n}{j} [\mathbb{P}(X \leq x)]^j [1 - \mathbb{P}(X \leq x)]^{n-j}.$$

In this section we will consider (static) thresholds of the form

$$T^k = \text{median}(X_{(k)})$$

for the optimal stopping problem. That is, the gambler sets the threshold T^k for every box and simply chooses the first box whose value exceeds this threshold. This is called the static threshold strategy based on T^k.

j) **[3 points]** Write a function that takes as input a probability distribution $\mathcal{F}$ (a MAKEDIST() object) of a random variable, and values n and k. It should output the median of the k-th order statistic $X_{(k)}$ of n I.I.D. random variables $X_1,\dots,X_n$ distributed according to $\mathcal{F}$, i.e., the threshold T^k.
Compute the median using FZERO() *with as initial guess the mean of* $\mathcal{F}$.

k) **[1 point]** Test your function from part j) on a Halfnormal distribution with $\mu = 1$ and $\sigma = 11$, for $n = 100$ and $k = 66$.

In the final question, you will set up your own numerical simulation framework to test which value of k yields the best result for the gambler when she uses the static threshold strategy based on T^k (as explained above). There is not necessarily one correct answer to this question; the most important thing here is that you justify your modelling choices (in the report).

ℓ) **[6 points]** Carry out the following three steps:

- ○ Choose three probability distributions that have different properties. Explain your choices in the report.
- ○ Design a numerical simulation framework to decide for which k the static threshold T^k yields the highest payoff for the stopping problem (as in the introduction) for an instance with n I.I.D. distributions.

[5] If you are not familiar with this concept, please have a look at some literature. The same holds for the concept of the median.

- Apply this framework to the three chosen probability distributions (being three possible choices for the common I.I.D. distribution in the previous point) and discuss your results.