Pakistan Journal of Statistics and Operation Research

A linear programming-based approach to estimate discrete probability functions with given quantiles

Zohre Nikooravesh[1*], Javad Tayyebi[2]

* Corresponding Author

1. Department of Basic Sciences, Birjand University of Technology, Birjand, Iran, nikooravesh@birjandut.ac.ir
2. Department of Industrial Engineering, Birjand University of Technology, Birjand, Iran, javadtayyebi@birjandut.ac.ir

Abstract

The aim of this paper is to estimate probability distribution functions with maximum entropy and known quantiles. The paper formulates the problem as a nonlinear optimization problem, and converts it into a system of nonlinear equations by Lagrange multipliers method. Finally, an efficient method is proposed to obtain a solution of the nonlinear system. The method needs to solve a linear programming problem in each iteration. Since linear programming problems can be solved in a reasonable time, our proposed method is faster than generic methods of solving nonlinear programming problems. Several computational experiment are provided to demonstrate the performance and validation of our proposed method.

Key Words: Maximum entropy problem; Nonlinear optimization; Lagrange multipliers method; Linear programming.

Mathematical Subject Classification: 46N10, 65K05, 94A17.

1. Introduction

In probability and statistics, a random variable is described informally as a variable whose values depend on outcomes of a random phenomenon. The formal mathematical definition of random variables is a topic in probability theory. A random variable is a measurable function defined on a probability space that maps from the sample space, the set of all outcomes of a random phenomenon, to the set of real numbers. Whenever the image, or range, of a random variable is countable, it called a discrete random variable. Moreover, a probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for a random variable.

Sometimes it is needed to find a probability distribution function by some initial observations. If some data are available, we achieve the goal by fitting a probability function on the data. There are a variety of methods, such as method of moments, maximum spacing estimation, method of L-moments (Hosking, 1990), maximum likelihood method (Aldrich et al., 1997), and statistical software (Oosterbaan, 2019) that can be used to obtain a probability distribution function. These methods may provide several different solutions for one problem because their approach is different.

In the case that some certain parameters of a distribution are known, instead of its data, there are also several methods to estimate the distribution. Since the selection of an appropriate distribution depends on several different factors, these methods usually do not provide a unique solution. For example, suppose that we know the data are symmetrically distributed around the mean while the frequency of occurrence of data farther away from the mean diminishes. One may select the normal distribution, the logistic distribution, or the student's t-distribution. The first two cases are very similar, while the last one, with one degree of freedom, has "heavier tails" meaning that values farther from the mean often occur relatively more, i.e., the kurtosis is higher.

By a pessimistic viewpoint, the maximum entropy provides an approach to estimate a unique solution to the problem (Cover and Thomas, 2006). Let us discuss it in the following. Shannon (1948) first introduced entropy for a random variable X with a sample space $S = \{x_1, x_2, \dots, x_n\}$ and corresponding probabilities $p_i, i = 1,2, \dots n.$ as follow:

$$H(X) = -\sum_{i=1}^{n} p_i log p_i,$$

where $\sum_{i=1}^{n} p_i = 1$and $p_i \geq 0$ for $i = 1,2, \dots n$. In information theory, the maximum entropy problem is formulated as follows (Cover and Thomas, 2006):

$$\max \quad H(X) \tag{1a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \tag{1b}$$

$$\sum_{i=1}^{n} p_i g_j(x_i) = \beta_j, \quad j = 1,2, \dots, n, \tag{1c}$$

$$p_i \geq 0, \quad i = 1,2, \dots, n. \tag{1d}$$

where β_j's are known real numbers and g_j's are real-value functions. The problem is to estimate a discrete probability function satisfying some given conditions (1c). Note that p_i is the decision variable of the problem, which indicates the probability of observing x_i. So its value must be non-negative (1d), and furthermore, the sum of its values must be equal to one (1b) to satisfy the principles of probability.

Let us now review some papers in this field. At first, Jaynes (1957) estimated a probability distribution with minimal probability under some certain conditions using Shannon's maximum entropy. Subsequently, many researchers focused on this issue. Zografos (2008) studied the features, applications and generalization of the maximum entropy problem. Landsman and Makov (1999) and Najafabadi et al. (2012) used it in belief theory. Krvavych and Mergel (2000) and Sachlas and Papaioannou (2014) examined the distribution modeling using the maximum entropy method. Most works were focused on estimating continuous probability functions.

Dai et al. (2016) introduced the concept of maximum entropy in the reliability context. In the case of the maximum probability of entropic discrete functions, Van der Straeten (2009) expressed the concept of maximum entropy in random processes, especially in Markov chains. Then, Chliamovitch et al. (2015) completed his work. Moreover, Basset (2015) summarized this issue. Many researchers have used maximum entropy method to find a probability function with highest uncertainty under some known conditions. For example, Zhao and Zhang (2011) proposed an ensemble neural network, which combines the component networks using the entropy theory. The entropy-based ensemble neural network searches the best structure of each component network first, and employs entropy as an automating design tool to determine the best combining weights. Bajgiran et al. (2020) explored the maximum entropy models that are the minimum elaborations of the uniform and moment-based ME models by quantiles. This property provided a diagnostic for the utility of elaboration in terms of the information value of each type of information over the other. They said "the maximum entropy model with quantiles and moments is represented as the mixture of truncated distributions on consecutive intervals whose shapes and existence are determined by the moments". In another article, the application of maximum entropy principle has been discussed to establish a probability distribution when the mentioned summary statistics are available, and its extension to moment constraints has been introduced to satisfy the requirements of metrology (Barzdajn, 2014).

In this paper, we investigate the maximum entropy problem of estimating discrete probability functions whenever some quantiles are known. Similar to the above argument, we obtain a nonlinear system and propose a novel method to solve the system. Our proposed method has an iterative process and needs to solve a linear programming problem in each iteration. Since there are several efficient methods, such as interior point methods (Vanderbei, 2015), to solve linear programming problems, our proposed method obtain a solution of the system in a reasonable time.

Although there are many numbers of papers to estimate probability functions for continuous random variables, to the best of our knowledge, there is not any paper to investigate our proposed case (for discrete random variables with known quantities). Specially, the approach of solving the obtained nonlinear optimization problem is not presented in any paper. This is a bad news, because we cannot compare our results with the other available approaches. To

resolve the problem, the validation of approach is evaluated by comparing its solutions with ones obtained from the nonlinear programming solver of Matlab software whose solutions are exact.
The reminder of this paper is organized as follows. Section 2 defines the problem and formulates it, mathematically. In Section 3, our proposed approach is presented to solve the problem. In Section 4, some computational experiments are conducted to demonstrate the validation and performance of the proposed approach. Finally, Section 5 points some concluding remarks, and proposes some subjects to future works.

2. Problem statement

Consider a discrete random variable X with a sample space $S = \{x_1, x_2, \dots, x_n\}$ in which $x_1 < x_2 < \cdots < x_n$. One can find many different distribution functions with some predetermined information, such as mean, variance, moments or quantiles. In this paper, we want to obtain a probability distribution function of the discrete random variable X with prescribed quantiles. By a pessimistic viewpoint, we select one that has the greatest entropy value among all distributions with this property. This function will have the highest possible uncertainty. Consequently, the actual probability distribution function will be more definite than this function. To formulate the problem, assume that α_i, $i = 1,2, \dots, r-1$, is the ith quantile with rank r such that

$$\alpha_1 < \alpha_2 < \cdots < \alpha_{r-1}.$$

Moreover, suppose that the value of $\alpha_j, j = 1,2, \dots, r-1$, belongs to the interval $[x_{i_j}, x_{i_j+1})$. So, we can formulate the problem of obtaining p_i's with prescribed quantiles α_j as follows:

$$\text{Max} \quad z = -\sum_{i=1}^{n} p_i log p_i, \tag{2a}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} p_i = 1, \tag{2b}$$

$$\sum_{i=1}^{i_j} \left(p_i + \frac{\alpha_j - x_{i_j}}{x_{i_j+1} - x_{i_j}} p_{i_j+1} \right) = \frac{j}{r}, \quad j = 1,2, \dots, r-1, \tag{2c}$$

$$p_i \geq 0, \quad i = 1,2, \dots, n. \tag{2d}$$

This is an optimization problem containing nonlinear objective function (2a) as well as linear constraints (2b-2d). So it is a constrained nonlinear programming problem. By multiplying the objective function by −1, one can convert the problem into a minimization problem. Similar to the papers (Arandjelovi´ c et al., 2014), (Van der Straeten, 2009), and (Templeman and Xingsi, 1987), we neglect the constraint $p_i \leq 1$ because this is satisfied by the constraints (2b) and (2d).

3. Our proposed method

In this section, we convert problem (2) into a nonlinear system. Then, we propose a method based on linear programming to solve this system.
To find an optimal solution of problem (2), we need to use nonlinear optimization techniques. One of the most popular techniques is Lagrange multipliers method. This method adds the constraints with some coefficients to the objective function. Hence, the problem changes to an unconstrained nonlinear programming problem. Then, by setting partial differentiations of the new objective function equal to zero, a critical point is found.
Now, consider an arbitrary instance of problem (2). To construct the Lagrange function, we do neglect nonnegativity constraint (2d), although this constraint will be added later to the problem. The corresponding Lagrangian function is

$$L(p, \lambda) = \sum_{i=1}^{n} p_i log p_i + \sum_{j=1}^{r-1} \lambda_j \left(\sum_{i=1}^{i_j} \left(p_i + \frac{\alpha_j - x_{i_j}}{x_{i_j+1} - x_{i_j}} p_{i_j+1} \right) - \frac{j}{r} \right) + \lambda_r \left(\sum_{i=1}^{n} p_i - 1 \right) \tag{3}$$

in which λ_j's are Lagrange multipliers.
Instead of solving problem (2), it is needed to find a non-negative solution that minimizes the function $L(p, \lambda)$. Since the entropy function is concave (For a proof, see (Cover and Thomas, 2006), Theorem 2.7.3) and constraints (2b-2d)

are linear, it follows that $L(p,\lambda)$ is convex. Hence, any local minimum point of $L(p,\lambda)$ is a global minimum. To find a local minimum, it is sufficient to look for a critical point. In the other words, we find a non-negative solution for the following system:

$$\frac{\partial L(p,\lambda)}{\partial p_i} = 0, \;\; i = 1,2,\dots,n, \tag{4}$$

$$\frac{\partial L(p,\lambda)}{\partial \lambda_j} = 0, \;\; j = 1,2,\dots,r. \tag{5}$$

By substituting (3) in (4) and (5), we have

$$log p_i + 1 + a_i^T \lambda = 0, \;\; i = 1,2,\dots n, \tag{6a}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{6b}$$

$$\sum_{i=1}^{i_j} \left(p_i + \frac{\alpha_j - x_{i_j}}{x_{i_j+1} - x_{i_j}} p_{i_j+1} \right) = \frac{j}{r}, \;\; j = 1,2,\dots,r-1, \tag{6c}$$

where $\lambda = [\lambda_1,\dots,\lambda_r]^T$ and a_i is the ith column of the coefficient matrix of the problem (2). If $i = i_{j_0}$, then its jth element of a_i is as

$$a_{ji} = \begin{cases} 0 & j \in \{1,2,\dots,j_0\}, \\ \dfrac{\alpha_j - x_i}{x_{i+1} - x_i} & j = j_0 + 1, \\ 1 & j \in \{j_0 + 2,\dots,r\}. \end{cases}$$

Otherwise, its jth element is as follows:

$$a_{ji} = \begin{cases} 0 & j \in \{1,2,\dots,j_0\}, \\ 1 & j \in \{j_0 + 1,\dots,r\}. \end{cases}$$

In system (6), only equation (6a) is nonlinear due to the existence of $log p_i$. Although one can use any method of nonlinear systems, such as Newton method (Deuflhard, 2011), solving the system, it is not an easy issue due to its nonlinearity.

To solve this system, we propose an iterative procedure. It generates a vector $p^{(k)}$ in kth iteration which is an estimation of p. The process begins with an initial vector $p^{(0)} > 0$. Suppose that the vector $p^{(k-1)}$ is found in $(k-1)^{\text{th}}$ iteration. The following problem is solved in kth iteration to generate $p^{(k)}$.

$$\min \quad w = \sum_{i=1}^{n} \left| p_i^{(k-1)} q_i - (log p_i^{(k-1)}) p_i \right|, \tag{7a}$$

$$\text{s.t.} \quad q_i + 1 + a_i^T \lambda = 0, \;\; i = 1,2,\dots n, \tag{7b}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{7c}$$

$$\sum_{i=1}^{i_j} \left(p_i + \frac{\alpha_j - x_{i_j}}{x_{i_j+1} - x_{i_j}} p_{i_j+1} \right) = \frac{j}{r}, \;\; j = 1,2,\dots,r-1, \tag{7d}$$

$$p_i \geq 0, \;\; i = 1,2,\dots,n. \tag{7e}$$

Constraints (7b-7d) are the same equations of system (6) with this difference that a new variable q_i is replaced with $log(p_i)$. Constraint (7e) is added to the problem to satisfy the nonnegativity of p_i. The purpose of solving the problem is to find the vectors p^*, $q^*$and λ^* so that the distance between $p^{(k-1)} q^*$and $p^* log p^{(k-1)}$ is minimized. So q^* is an appropriate estimation of $log p_i$. After solving the problem, we set $p^{(k)} = p^*$ and repeat the process until the optimal objective value w^* becomes zero.

It is remarkable that the condition $p_i > 0$ is required to be added to the problem because the logarithm function is not defined at zero. However, we do not need to express explicitly it in practice because $log(0)$ is defined to be negative infinity, as a limitation value, in many programming languages. Hence, if this value appears in the

objective function, it increases the objective value to positive infinity. This never occurs because it is a minimizing optimization problem.

Theorem 3.1. *If problem (7) has an optimal solution p^*, q^* and λ^* with the optimal value $w^* = 0$, then the vectors p^* and λ^* are a solution of system (6).*

Proof. The equality $w^* = 0$ guarantees that $q_i^* = \log(p_i^{(k-1)})$ and $p_i^* = p_i^{(k-1)}$ for every $i = 1,2,\ldots,n$. So λ^* and p^* are a solution from system (6). □

Because of the existence of absolute value, problem (7) is a nonlinear optimization problem in the current form. If we define non-negative variables x_i^- and x_i^+ for which $x_i^+ - x_i^- = p_i^{(k-1)} q_i - \left(log p_i^{(k-1)}\right) p_i$, then the problem is converted to a linear programming problem. Using this linearization technique, problem (7) is rewritten as follows:

$$\min \quad w = \sum_{i=1}^{n} x_i^- + x_i^+, \tag{8a}$$

$$\text{s.t.} \quad x_i^+ - x_i^- = p_i^{(k-1)} q_i - \left(log p_i^{(k-1)}\right) p_i, \tag{8b}$$

$$q_i + 1 + a_i^T \lambda = 0, \quad i = 1,2,\ldots n, \tag{8c}$$

$$\sum_{i=1}^{n} p_i = 1, \tag{8d}$$

$$\sum_{i=1}^{i_j} \left(p_i + \frac{\alpha_j - x_{i_j}}{x_{i_j+1} - x_{i_j}} p_{i_j+1} \right) = \frac{j}{r}, \quad j = 1,2,\ldots,r-1, \tag{8e}$$

$$p_i, x_i^+, x_i^- \geq 0, \quad i = 1,2,\ldots,n. \tag{8f}$$

The reason for transforming the nonlinear optimization problem to a linear programming problem is that linear programming problems can be solved exactly in a finite number of iterations by several algorithms, such as active set, and interior point methods (Terlaky, 2013). This aids us to use the inherent simplicity of linear programming.
Now, we are ready to explain our proposed approach in complete details. This approach begins with an initial solution $p^{(0)} > 0$. This solution can be obtained from solving the system of linear equations (8d-8f). Then, problem (8) is solved for every number $k = 1,2,3,\ldots$ and the sequence $\{p^{(k)}\}$ is generated. The process is repeated until the optimal value becomes zero (see Theorem 3.1).

4. Computational results

In this section, some computational experiments are conducted to demonstrate the validation and the performance of our proposed approach. All experiments are performed on a 4-core computer with 8GB of RAM and Windows 10 operating system. We used the software of MATLAB for implementing programs.
To perform computations, two stopping conditions are used. The first condition is $w^* > 10^{-10}$, that is, the process terminates if the optimal value of problem (8) is sufficiently close to zero (see theorem 3.1). Results show that the accuracy of solutions is guaranteed even if the number of iterations is small. For this reason, the second stopping condition is that the number of iterations is at most equal to 5.
To solve linear programming problems, the function "*linprog*" of Matlab is used. We compared our proposed approach with a nonlinear optimization method, called the intrinsic method that runs directly on problem (2). This approach is the default method used in the command "*fmincon*" of Matlab.

4.1. Validation

To check the validation of the approach, we have solved a small instance of the problem and have compared its solution with the one obtained from solving the problem by Matlab solver. For this purpose, we have supposed that X is a random variable on a sample space $S = \{1,2,\ldots,20\}$, moreover,

$$\alpha_1 = 2, \alpha_2 = 5.5, \alpha_3 = 11, \alpha_4 = 12.5,$$

with rank r = 5.
Our approach solved the problem in 0.425796 seconds while Matlab solved it in 0.706412 seconds. The entropy values are respectively −2.7966 and −2.7967. Table 1 shows the results and Figure 1 depicts them. It is obvious that the solutions are very close together. The results show the validation of our approach for this small instance.

However, comparing entropy values of randomly generated large instances is also another useful tool to check the validation of our approach. This issue together with comparing running times are performed in the next subsection.

Table 1: Comparing our approach's solution with the exact solution obtained from Matlab solver

X	1	2	3	4	5	6	7	8	9	10
Exact sol.	0.1	0.1	0.0591	0.0591	0.0591	0.0457	0.0354	0.0354	0.0354	0.0354
Our sol.	0.1	0.1	0.059	0.059	0.059	0.0457	0.0354	0.0354	0.0354	0.0354
X	11	12	13	14	15	16	17	18	19	20
Exact sol.	0.0354	0.1673	0.0655	0.0239	0.0239	0.0239	0.0239	0.0239	0.0239	0.0239
Our sol.	0.0354	0.1682	0.0636	0.024	0.024	0.024	0.024	0.024	0.024	0.024

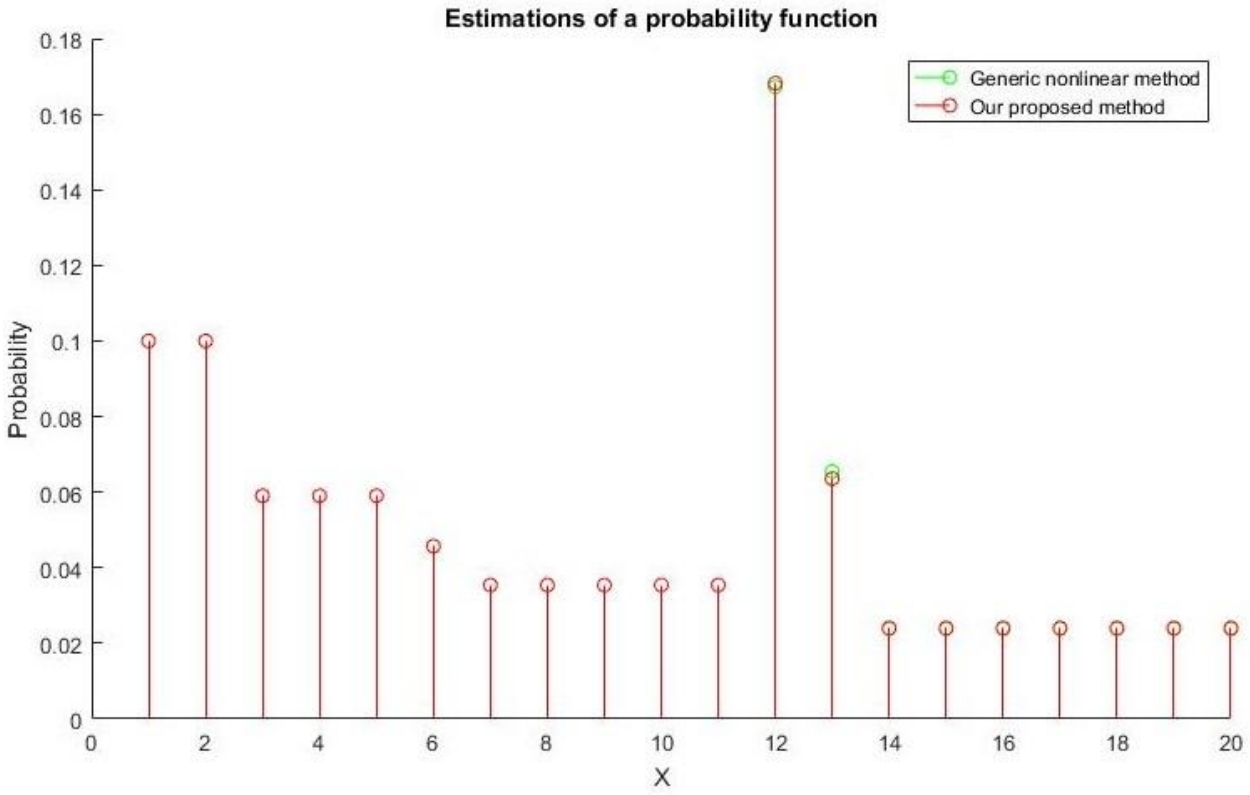

Figure 1 Comparing our approach's solution with the exact one

4.2. Performance

Now we perform several randomly generated experiments to guarantee the performance of the method. In experiments, only the parameters n and r are determined exactly, and other data are generated randomly from the interval $[0, n]$ by uniform distribution, i.e.,

$$x_i \sim U(0, n), \qquad i = 1, \dots n,$$
$$\alpha_j \sim U(0, n), \qquad j = 1, \dots, r.$$

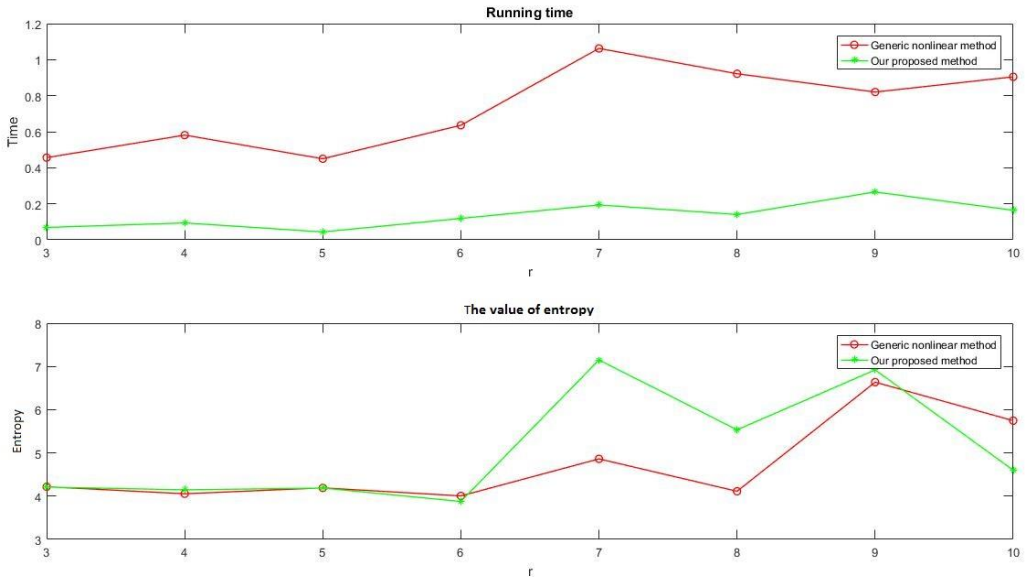

Figure 2: The performance graph for $n = 100$ and different values r

All experiments were repeated ten times and the average of results (running time and entropy value) were reported. Two types of experiments were performed for comparison. In the first experiment, it is assumed that n = 100 and the number of quantiles varies from $r = 3$ to $r = 10$ (see Figure 2). In the second experiment, the number of quantiles is equal to 4, and $n = 50,75,100,125,150,175,200$ (see Figure 3).
As we expected, the running time of our proposed approach was significantly less than that of the generic nonlinear method because ours applies linear programming methods as fast tools to solve the nonlinear problem. On the other hand, the entropy values obtained from both the approaches were approximately the same. This establishes that our proposed approach has a fairly good performance both in running time and in solution accuracy.

4.3. Sensitivity Analysis

Here, we conduct a sensitivity analysis of our approach. Since the problem does not contain any single parameter, we run the sensitivity analysis by changing the range of randomly generated data. Recall that input data were generated randomly as

$$x_i \sim U(0, n), \qquad i = 1, \dots n,$$
$$\alpha_j \sim U\left(0, \frac{1}{s}n\right), \qquad j = 1, \dots, r.$$

for $s = 1$. Now, assume that $n = 20$ and s varies from 1 to 4. This sequentially changes the range of α_j. Specially, all values α_j are in the range $[0,5]$ for $k = 4$ to show the probability distribution function has positive skewness. Table 2 and Figure 4 provide the numerical results as well as the graphs obtained from the sensitivity analysis. The graphs show that the skewness changes significantly whenever s varies from 1 to 4. By comparing the numerical results, it is easily observed that our approach performs correctly and its time is better than Matlab Solver.

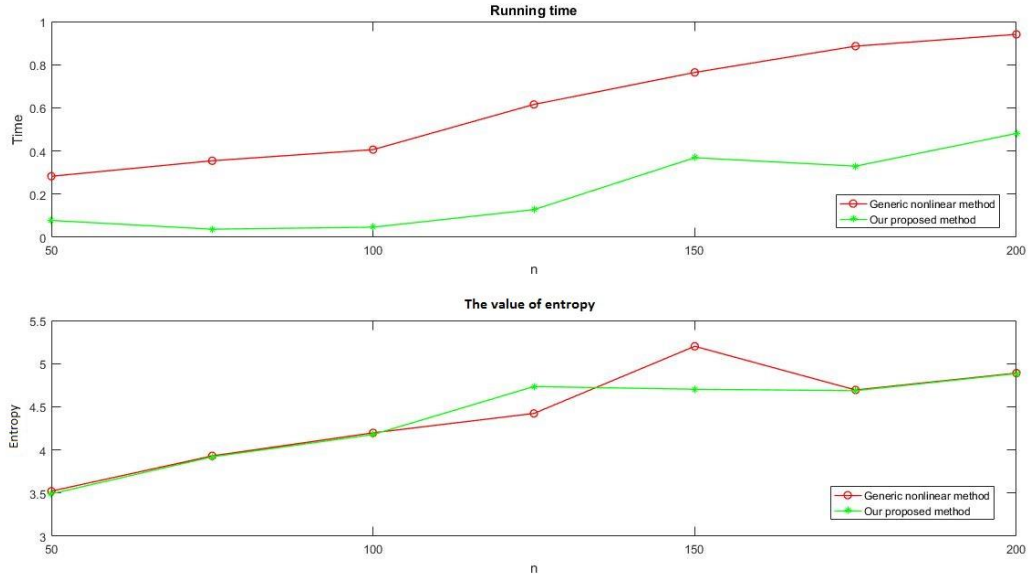

Figure 3: The performance graph for $r = 4$ and different values n

5. Conclusions

In this paper, the problem of estimating a discrete probability function with given quantiles was investigated. The well-known Lagrange multipliers method was used to obtain a nonlinear system for finding a probability distribution with some known quantiles and maximum entropy value. Then, a novel approach was proposed to solve the nonlinear system.

The use of linear programming approach is the most important feature of our approach because in spite of the fact that the problem is inherently nonlinear, we could design the approach that uses the intrinsic simplicity of linear programming problems at each iteration. This causes that the approach has a fairly good performance both in running time and in solution accuracy. Another feature of the approach is that the linear programming problems corresponding to two consecutive iterations are different only in a set of constraints. So one can apply the sensitivity analysis approach of linear programming problems to obtain the optimal solution of an iteration from that of the previous iteration.

The problem has the property that any probability value p_i cannot be zero because the logarithmic value at zero is assumed to be equal to $-\infty$. This imposes the limitation of using the well-known simplex method because this method obtains a basic optimal solution which contains many nonbasic variables being equal to zero. So other methods have to be used for solving linear programming problems.

As an application of our approach, it will be meaningful to investigate the maximum entropy problem for estimating continuous probability functions whenever some quantiles are known. For this purpose, one can apply a numerical discretization method and then, uses our approach to obtain a near-optimal solution. Due to the existence of different discretization methods and time-consuming computations, this can be a vital issue in this field. As another suggestion, one can use nonlinear programming techniques for designing other approaches of estimating probability distribution functions.

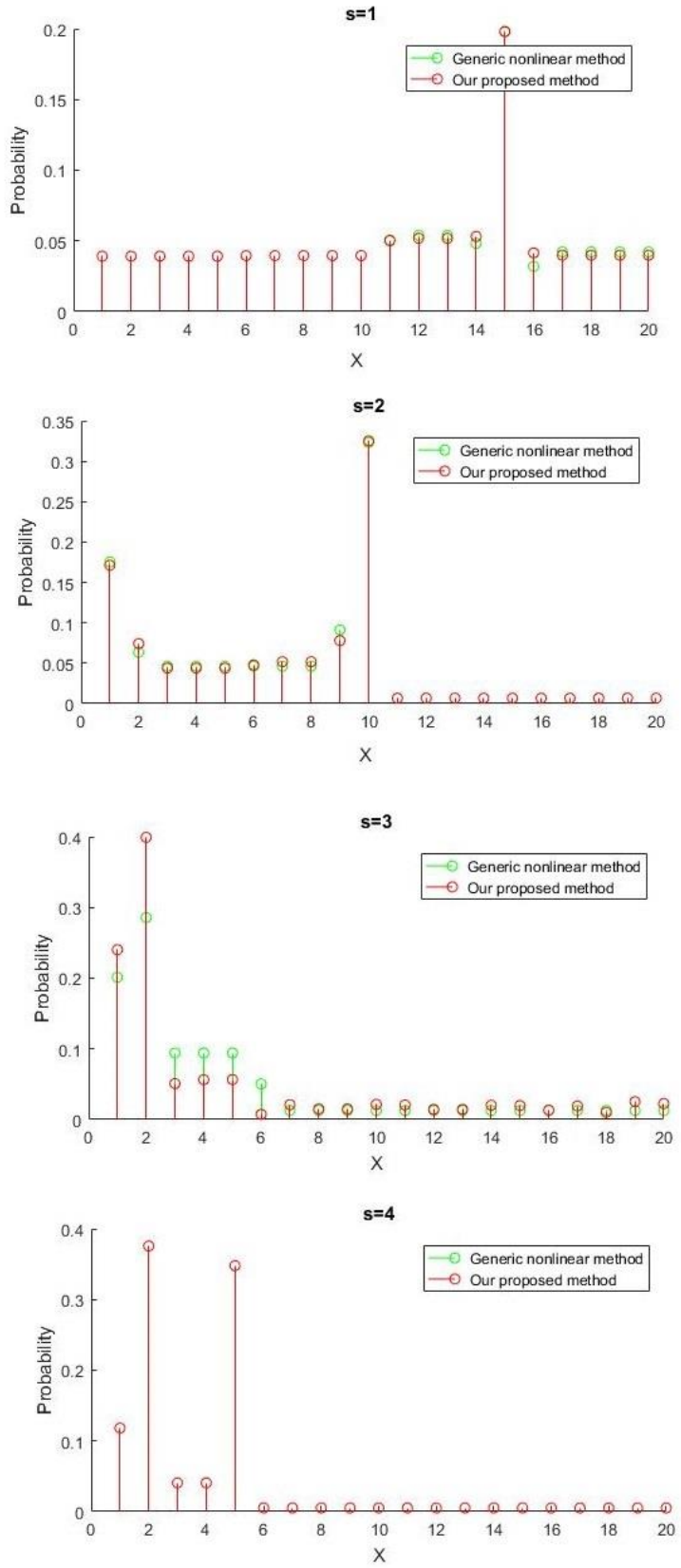

Figure 4: The graphs of probability distribution in the sensitivity analysis

Table 2: The numerical results of sensitivity analysis

S		Time	Entr.	1	2	3	4	5	6	7	8	9
1	Sol.	0.404	2.79	0.090	0.090	0.030	0.024	0.024	0.024	0.024	0.024	0.024
	Ours	0.016	2.8	0.085	0.085	0.050	0.023	0.023	0.023	0.023	0.023	0.023
2	Sol.	0.125	2.18	0.253	0.060	0.074	0.033	0.258	0.175	0.010	0.010	0.010
	Ours	0.061	2.18	0.234	0.038	0.09	0.100	0.197	0.202	0.009	0.009	0.009
3	Sol.	0.078	2.32	0.183	0.206	0.140	0.133	0.130	0.019	0.013	0.013	0.013
	Ours	0.0313	2.32	0.183	0.206	0.140	0.132	0.129	0.025	0.013	0.013	0.013
4	Sol.	0.140	3.73	0.359	0.499	0.00	0.856	0.066	0.066	0.066	0.066	0.066
	Ours	0.0625	3.64	0.330	0.459	0.00	0.786	0.061	0.061	0.061	0.061	0.061
s		10	11	12	13	14	15	16	17	18	19	20
1	Sol.	0.024	0.027	0.041	0.041	0.041	0.041	0.047	0.066	0.066	0.070	0.178
	Ours	0.023	0.028	0.040	0.040	0.040	0.040	0.047	0.061	0.061	0.085	0.172
2	Sol.	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010	0.010
	Ours	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009	0.009
3	Sol.	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
	Ours	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013	0.013
4	Sol.	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066	0.066
	Ours	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061	0.061

References

1. Aldrich, J. et al. (1997). Ra fisher and the making of maximum likelihood 1912-1922. Statistical science,12(3):162–176.
2. Arandjelovi´ c, O., Pham, D.-S., and Venkatesh, S. (2014). Two maximum entropy-based algorithms for running quantile estimation in nonstationary data streams. IEEE Transactions on circuits and systems for video technology, 25(9):1469–1479.
3. Bajgiran, A. H., Mardikoraem, M., and Soofi, E. S. (2020). Maximum entropy distributions with quantile information. European journal of operational research.
4. Barzdajn, B. (2014). Maximum entropy distribution under moments and quantiles constraints. Measurement,
5. 57:102–107
6. Basset, N.(2015). A maximal entropy stochastic process for a timed automaton. Information and Computation, 243:50–74.
7. Chliamovitch, G., Dupuis, A., and Chopard, B. (2015). Maximum entropy rate reconstruction of markov dynamics. Entropy, 17(6):3738–3751.
8. Cover, T. and Thomas, J. (2006). Elements of Information Theory, (2nd edn, 2006).
9. Dai, H., Zhang, H., and Wang, W. (2016). A new maximum entropy-based importance sampling for reliability analysis. Structural Safety, 63:71–80.
10. Deuflhard, P. (2011). Newton methods for nonlinear problems: affine invariance and adaptive algorithms, volume 35. Springer Science & Business Media.
11. Hosking, J. R. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. Journal of the Royal Statistical Society: Series B (Methodological), 52(1):105–124.
12. Jaynes, E. T. (1957). Information theory and statistical mechanics. ii. Physical review, 108(2):171.
13. Krvavych, Y. and Mergel, V. (2000). Large loss distributions: probabilistic properties, evt tools, maximum entropy characterization. In Proceedings of the 31st ASTIN Colloquium, Sardinia, Italy.
14. Landsman, Z. and Makov, U. E. (1999). Credibility evaluation for the exponential dispersion family. Insurance: Mathematics and Economics, 24(1-2):23–29.
15. Najafabadi, A. T. P., Hatami, H., and Najafabadi, M. O. (2012). A maximum-entropy approach to the linear credibility formula. Insurance: Mathematics and Economics, 51(1):216–221.

16. Oosterbaan, R. (2019). Software for generalized and composite probability distributions. International Journal of Mathematical and Computational Methods, 4.
17. Sachlas, A. and Papaioannou, T. (2014). Residual and past entropy in actuarial science and survival models. Methodology and Computing in Applied Probability, 16(1):79–99.
18. Shannon, C. E. (1948). A mathematical theory of communication. The Bell system technical journal, 27(3):379–423.
19. Templeman, A. B. and Xingsi, L. (1987). A maximum entropy approach to constrained non-linear programming. Engineering Optimization+ A35, 12(3):191–205.
20. Terlaky, T. (2013). Interior point methods of mathematical programming, volume 5. Springer Science & Business Media.
21. Van der Straeten, E. (2009). Maximum entropy estimation of transition probabilities of reversible markov chains. Entropy, 11(4):867–887.
22. Zhao, Z. and Zhang, Y. (2011). Design of ensemble neural network using entropy theory. Advances in Engineering Software, 42(10):838–845.
23. Zografos, K. (2008). On some entropy and divergence type measures of variability and dependence for mixed continuous and discrete variables. Journal of statistical planning and inference, 138(12):3899–3914.