Introduction to Nonlinear Programming

Term Project

Lifu Chen[1] and Mingyuan Liu[2]

[1] Graduate Institute of Communication Engineering
r09942043@ntu.edu.tw
[2] Institute of Industrial Engineering
r05546018@ntu.edu.tw

Parameter estimation plays a crucial role in statistics, communication systems, machine learning, and other domains. There are many techniques for parameter estimation. A common framework used is maximum likelihood estimation. However, while maximum likelihood estimators enjoy good properties in statistics, such as consistency and asymptotic efficiency, they often have no closed-form expression due to the necessity of solving nonlinear optimization problems. In this term project, we applied the methods for nonlinear programming to solve the maximum likelihood estimate for fitting the score distribution of the students taking EE 2007.

1 Introduction

A statistical model is a collection of probability distributions, denoted by $M(\theta) = \{f_\theta \mid \theta \in \Theta\}$, where θ are the parameters that encode important properties about the model. We often propose a statistical model to explain some observed phenomenon. Since θ contain vital information about the model, they are useful in the description of the stochastic structure with respect to the data. Therefore, how to find the best parameter choice with observed data is in our interest.
There are many different techniques for parameter estimation. One popular tool is maximum likelihood estimation. Maximum likelihood estimation enjoys nice theoretical properties in statistics and can be applied to a great variety of statistical problems, such as classification and model selection. Its general utility is one of the major reasons for its importance in parameter estimation.
Suppose that i.i.d. random variables X_i have a common density function $f(x \mid \theta)$. Given observed values $X_i = x_i$, where $i = 1, \ldots, n$, the likelihood function of θ with respect to the observed data $\{x_1, x_2, \ldots, x_n\}$ is defined as

$$lik(\theta) = f(x_1, x_2, \ldots, x_n|\theta) = \prod_{i=1}^{n} f(x_i \mid \theta)$$

The maximum likelihood estimate of θ is that value of θ that maximizes the likelihood. That is, the maximum likelihood estimator of θ is defined as

$$\hat{\theta}(X_1, \ldots, X_n) = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(X_i \mid \theta)$$

However, in most practical models such as gamma and beta models, maximum likelihood estimators have no closed-form solutions. Thus, to know the maximum likelihood estimate, we need the help of nonlinear optimization methods.

2 Model Formulation

We collected the midterm scores from the students taking EE 2007 and wondered if any common distributions fit their score distribution well. To answer the question, we set up three different statistical models for the grades and applied maximum likelihood estimation to find the best parameter choice for each model.

2.1 Model 1: Normal

The first model is under the assumption that students' scores follow a normal distribution. Therefore, our statistical model is

$$M(\mu,\sigma) = \{f_{\mu,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \mid \mu \in \mathbf{R},\ \sigma > 0\}$$

Given the scores of the students $\{x_1,\ldots,x_n\}$, its maximum likelihood estimate is

$$\hat{\theta} = \underset{\mu\in\mathbf{R},\ \sigma>0}{\arg\max} \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2)$$

2.2 Model 2: Gamma

The second model is under the assumption that students' scores follow a gamma distribution. Therefore, our statistical model is

$$M(k,b) = \{f_{k,b}(x) = \frac{1}{\Gamma(k)b^k}x^{k-1}e^{-\frac{x}{b}},\ x \in (0,\infty) \mid k > 0,\ b > 0\}$$

Given the scores of the students $\{x_1,\ldots,x_n\}$, its maximum likelihood estimate is

$$\hat{\theta} = \underset{k>0,\ b>0}{\arg\max} \prod_{i=1}^{n} \frac{1}{\Gamma(k)b^k}x_i^{k-1}e^{-\frac{x_i}{b}}$$

2.3 Model 3: Beta

The third model is under the assumption that students' scores follow a scaled Beta distribution. That is, $X_i = 100B_i$, where B_i follows a Beta distribution. Therefore, our statistical model is

$$M(a,b) = \{f_{a,b}(x) = \frac{1}{100B(a,b)}(\frac{x}{100})^{a-1}(1-\frac{x}{100})^{b-1},\ x \in (0,100) \mid a > 0,\ b > 0\}$$

Given the scores of the students $\{x_1, \ldots, x_n\}$, its maximum likelihood estimate is

$$\hat{\theta} = \arg\max_{a>0,\ b>0} \prod_{i=1}^{n} \frac{1}{100B(a,b)} \left(\frac{x_i}{100}\right)^{a-1} \left(1 - \frac{x_i}{100}\right)^{b-1}$$

3 Solution Approach

Since the logarithmic function is monotonic, we have

$$\begin{aligned} \hat{\theta}(x_1, \ldots, x_n) &= \arg\max_{\theta \in \Theta} \prod_{i=1}^{n} f(x_i \mid \theta) = \arg\max_{\theta \in \Theta} \log\left(\prod_{i=1}^{n} f(x_i \mid \theta)\right) \\ &= \arg\max_{\theta \in \Theta} \sum_{i=1}^{n} \log(f(x_i \mid \theta)) \end{aligned}$$

Then it suffices to maximize $\sum_{i=1}^{n} \log(f(x_i \mid \theta))$ over the parameter space Θ. This modified optimization problem is more numerically stable than the original one since the likelihood function can be extremely small after many terms are multiplied. Thus, we decided to maximize the log-likelihood function instead of the likelihood function.
We used two methods to solve the optimization problem. One is the method of simulated annealing. The other is the method of Davidon-Fletcher-Powell.

3.1 Simulated Annealing

Simulated annealing (SA) is a random-search method for approximating the global optimum of a given function. It is useful in finding the global optimum in the presence of a lot of local optima. For problems where finding an approximate global optimum is more important than finding a precise local optimum, simulated annealing may be chosen instead of exact algorithms such as steepest descent. Therefore, it is useful in maximum likelihood estimation.
The algorithm of the Metropolis simulated annealing for finding the maximum likelihood estimate is described as follows:

1. Set up the initial temperature k and the number of iterations. Pick an initial point in the parameter space Θ and evaluate the log-likelihood at that point.
2. Repeat the following steps until the required number of iterations is reached:
3. Pick a new point at random near the old point and compute the log-likelihood
4. If the new value is better, accept it and go to Step 3.
5. If the new value is worse, then pick a random number between 0 and 1. Accept the new (worse) value if the random number is less than exp(change in log-likelihood/k). Otherwise, go back to the previous value

6. Periodically (e.g. every 100 iterations) lower the value of k (e.g. k <- 0.8k) to make it harder to accept bad moves.

3.2 Davidon-Fletcher-Powell

The Davidon-Fletcher-Powell (DFP) method is one of the quasi-Newton methods. When the log-likelihood function has only one local optimum, we can apply it to find the maximum likelihood estimate. That is, take the log-likelihood function as the objective function and run the algorithm we have learned in Lecture 6.

4 Simulation Result

4.1 Model 1: Normal

In the simulated annealing method, the temperature k is 50 and the number of iterations is 10000.
In the DFP method, the termination tolerance ϵ is 0.001.
The initial point for both methods is (μ, σ)=(50, 10).

Table 1: Maximum likelihood estimation for the normal model

Method	$(\hat{\mu}, \hat{\sigma})$	log-likelihood
SA	(48.01661, 15.67771)	-971.8942
DFP	(48.01667, 15.67809)	-971.8942

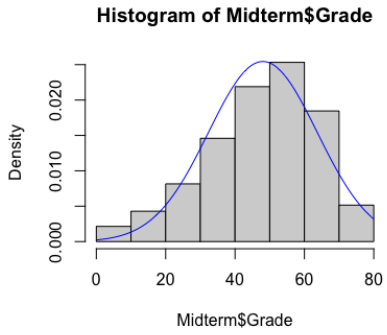


Figure 1: The estimated normal distribution on the histogram

4.2 Model 2: Gamma

In the simulated annealing method, the temperature k is 50 and the number of iterations is 10000.
In the DFP method, the termination tolerance ϵ is 0.001.
The initial point for both methods is (k, b)=(2, 3).

Table 2: Maximum likelihood estimation for the gamma model

Method	$(\hat{k}, \hat{b})$	log-likelihood
SA	(6.566189, 7.31295)	-1001.145
DFP	(6.566575, 7.312367)	-1001.145

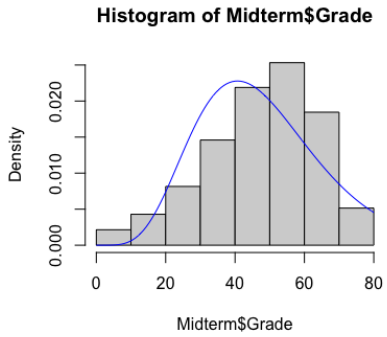


Figure 2: The estimated gamma distribution on the histogram

4.3 Model 3: Beta

In the simulated annealing method, the temperature k is 50 and the number of iterations is 10000.
In the DFP method, the termination tolerance ϵ is 0.001.
The initial point for both methods is (a, b)=(2, 3).

Table 3: Maximum likelihood estimation for the beta model

Method	$(\hat{a}, \hat{b})$	log-likelihood
SA	(4.091616, 4.516377)	-975.7882
DFP	(4.091402, 4.516018)	-975.7882

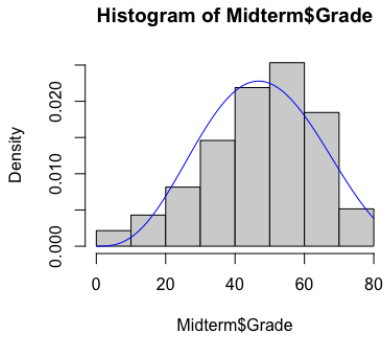


Figure 3: The estimated beta distribution on the histogram

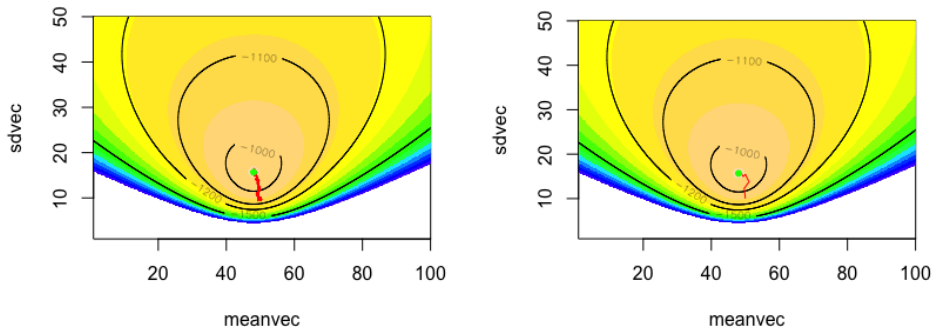
5 Discussion and Analysis

5.1 Simulated Annealing

The simulated annealing method mimics the annealing process in metallurgy, in which the temperature is carefully controlled to decrease slowly so that the material reaches a state with minimum free energy. The analogy is shown between the slow cooling in the annealing process and the slow decrease in the probability of accepting worse parameters in the algorithm. We can also see that in the early stage of its search when the temperature is high, it is more probable to accept worse parameters and thus to get out of the local optima, which enables the SA method to tackle the situation in which many local optima exist.

5.2 Convergence Speed

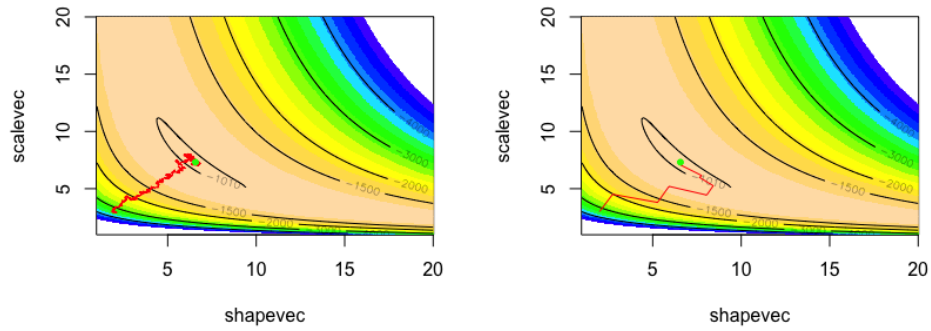
The simulated annealing method can deal with arbitrary objective functions and statistically guarantees to find an optimal solution. However, it cannot tell whether it has found the optimal solution and doesn't take the information about the system to be optimized into consideration. The price it pays is the long running time and reduced efficiency. In comparison, although the method of Davidon-Fletcher-Powell cannot guarantee the solution it finds is a global optimum, it converges to a local optimum quickly. Therefore, if we have more information about the optimization problem such as convexity or smoothness, the method of Davidon-Fletcher-Powell is preferred.
The difference of the convergence speed between the two methods can be visualized in the following figures:



(a) SA (8718 steps)

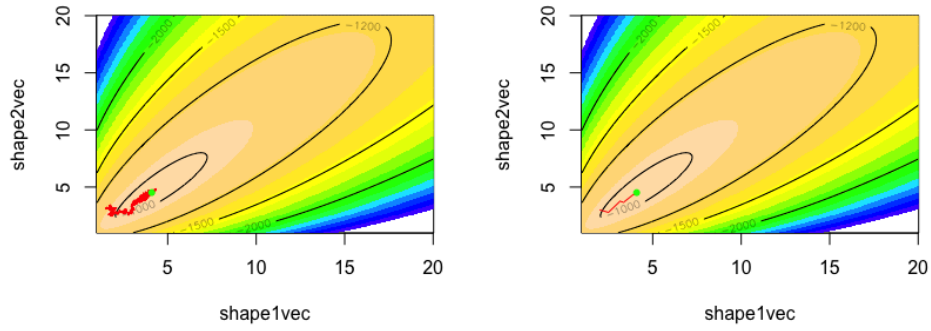
(b) DFP (10 steps)

Figure 4: Normal Model: The track of the iterations



(a) SA (8341 steps) **(b)** DFP (14 steps)

Figure 5: Gamma Model: The track of the iterations



(a) SA (9787 steps) **(b)** DFP (8 steps)

Figure 6: Beta Model: The track of the iterations

5.3 Goodness of Fit

To compare the goodness of fit of the three different models to our score distribution, we applied the Akaike information criterion (AIC). The Akaike information criterion is an estimator of prediction error and can be used to compare relative quality of statistical models for a given set of data. Its value is defined as

$$AIC = 2k - \log(like(\hat{\theta})),$$

where k is the number of the estimated parameters in a model and $\log(like(\hat{\theta}))$ is the maximum value of the log-likelihood function in the model.
The model with the distribution having the smallest AIC value is usually preferred.

Therefore, based on **Table 4**, the normal distribution fits our score distribution the best among the three models.

Table 4: AIC values for different models

Model	AIC
Normal	1947.78
Gamma	2006.29
Beta	1955.56

Finally, it is worth mentioning that in the first model under the assumption that the scores follow a normal distribution, the maximum likelihood estimator has a closed-form expression

$$\hat{\mu}_{MLE} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\hat{\sigma}_{MLE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu}_{MLE})^2}.$$

Thus, the maximum likelihood estimate is $\hat{\mu} = 48.01716$ and $\hat{\sigma} = 15.6783$, which agree with the simulation result in **Table 1**.

6 Conclusion

We applied two optimization methods to solve the maximum likelihood estimation problem in the parameter estimation. The performances of both optimization methods are accurate enough to find the best fitting distribution to the score distribution among popular distributions. When we know little about the system to be optimized, we would prefer the method of simulated annealing so that a global solution is statistically guaranteed. However, when we know more information, such as smoothness or convexity, about the system to be optimized, we would prefer the method of Davidon-Fletcher-Powell for its efficient convergence. After estimating the parameters, we compared the goodness of fit by the three different models and found that the normal distribution fits our score distribution the best.

7 Reference

1. Rice, John A. (2007), *Mathematical Statistics and Data Analysis*, 3rd Edition. Duxbury Press.

2. Bazaraa, Sherali, and Shetty. (2006), *Nonlinear Programming: Theory and Algorithms*, 3rd Edition. Wiley-Interscience.

3. Wikipedia contributors. "Simulated annealing." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 13 Apr. 2021. Web. 9 Jun. 2021.

4. https://kevintshoemaker.github.io/NRES-746/

5. http://www.new-npac.org/projects/cdroms/cewes-1999-06-vol1/cps615course/

6. https://www.spcforexcel.com/knowledge/basic-statistics/deciding-which-distribution-fits-your-data-best