2014 International Conference on Statistics and Mathematics (ICSM 2014)

Comparing Quasi Newton BFGS and Nelder Mead algorithm for Box-Cox Transformation

Wayan Permana Saputra, S.S.T.[a], Dr. Margaretha Ari Anggorowati[b],a*b*

[a]First affiliation, Jl. Otto Iskandardinata No.64C, Jakarta Timur 13330, Indonesia
[b]Second affiliation, Jl. Otto Iskandardinata No. 64C, Jakarta Timur 13330, Indonesia

Abstract

In parametric statistics test, there are some assumptions which must be fulfilled in order to make a valid conclusion. Some common assumptions in parametric statistics test are homoscedasticity or homogeneity of variance and assumption of normality. Many of us in social sciences deal with data that do not conform to assumptions of normality. To analyze non normal data with statistical test that require assumptions of normality, data should be transformed to a normal distribution. Box-Cox transformation is well known method and popular among statistician to get the normality (univariate and multivariate normal) because it doesn't require knowledge about data characteristics or trial and error to transform any data. Box-Cox method use estimation of parameter λ to do transformation, λ is obtained by maximize the Box-Cox function. Numerical method like quasi newton BFGS is commonly used to maximize the Box-Cox function because this method has fast convergence property but it may fail to convergence in some circumstance. In other hand, direct optimization algorithm like Nelder Mead may still be convergence in case which quasi newton BFGS fail to convergence but this algorithm generally has slower convergence property. In this paper, we will compare efficiency in term of function evaluation and processing time of the two algorithms to maximize Box-Cox function on some scenario.

Keywords: Box-Cox Transformation; Nelder Mead; Quasi Newton BFGS; Univariate Normal; Multivariate Normal

1. Introduction

Normality assumption has been a problem to many researcher in case using classical statistical method dealing with real data. Classical statistical method use normality distribution as basis for determining rejection region for hypothetical testing. Violation with these assumption might tend to directing us to the

* Corresponding author. Tel.: +6281317517895, +6287843987111; fax: (021) 8197577.
E-mail address: 10.6482@stis.ac.id, m.ari@stis.ac.id .

wrong decision. Given the importance of normal distribution, some option that we can choose in case violation against normality assumption according to Graybill(1976) in Sakia(1996) are:

- Ignore the violation of the assumptions and proceed with the analysis as if all assumption are satisfied.
- Decide what is the correct assumption in place, of the one that is violated and use a valid procedure that takes into account the new assumption.
- Design a new model that has important aspects of the original model and satisfies all the assumptions example by applying a proper transformation to the data or filtering out some suspect data point (outlying).
- Use a distribution free procedure that is valid even if various assumption are violated.

Among options available, transformation data to near normality is mostly used.

Transformation data to near normality is re-expressing data to another unit with the aim of obtaining data that meets normal distribution. According to Zimmerman(1998) and Osborne(2010), normal transformation not only beneficial to parametric statistical test but also can improve the accuracy of non parametric test. However the use of normal transformation must be careful. Transformation can change the nature of data and might be more difficult to interpret the result of transformed data.

There are some kind of transformation method. Osborne (2002) mentioned square root, log and inverse method. Those transformation requires researcher's knowledge about characteristic data to be able choosing the appropriate transformation method applied to data. That limitation often make researcher tried every method when not sure about characteristic data. Box and cox (1964) proposed transformation method which search optimum value from transformation Box-Cox function. The result is the best rank to perform data transformation, therefore Box-Cox transformation already included family of power transformation (square root, log, inverse, etc.).

Box-cox's transformation formula is available to univariate and multivariate data developed by maximum likelihood estimator function. Box-cox function can be solve using analytic approach where maximum value of function obtained by finding point that has differentiate equal to zero. However, that approach is manually difficult to use and require long time searches. In this case, optimization using numerical method is the appropriate solution. Numerical method which is generally known in MLE's parameter and standard error estimation are Newton Rapshon and quasi newton BFGS. Both of them become very popular because of small number iteration to reach convergency. To get convergent, both Newton Raphson and quasi newton BFGS need additional information in the form of gradient function each iteration to expect the next optimum alleged point, so both called gradient based optimization. Gradient based optimization algorithm has disadvantages in case of discontinuous functions, uncountable derivative value, etc. which make both method failed to convergent. Therefore we need non gradient based optimization which is more reliable to reach convergency, one that famous is Nelder Mead Simplex Direct Search. In this paper, we will compare efficiency of quasi newton BFGS and Nelder Mead when applied to Box-Cox transformation. Section 2.1 we will discuss about Box cox transformation, section 2.2 about Nelder Mead Simplex Direct Search, section 2.3 about quasi newton BFGS, and section 3 about data used and simulation process during research.

2. Methods

2.1. Box-Cox Transformation

In 1957 Tukey introduce a family of power transformation. The transformation equation expressed as :

$$y^{(\lambda)} = \begin{cases} y^{\lambda} & \lambda \neq 0 \end{cases} \qquad (1)$$

However, Tukey’s transformation didn’t take account of discontinuity at $\lambda=0$. Furthermore Box & Cox (1964) proposed a modified version of this family of power transformation. With regards in discontinuity at $\lambda=0$ the modified version transformation suggested by Box & Cox is

$$y^{(\lambda)}=\begin{cases}\frac{y^{\lambda}-1}{\lambda} & \lambda\neq 0\end{cases} \qquad (2)$$

And

$$y^{(\lambda)}=\begin{cases}\frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1} & \lambda_1\neq 0\end{cases} \qquad (3)$$

Estimation of λ can be done by Bayesian methods, by maximum likelihood based method, or by another method. Originally Box & Cox (1964) provide maximum likelihood method as well as Bayesian method for the estimation of the parameter λ. In this paper we will discuss about λ estimation using maximum likelihood method. As we know Box-Cox transformation is primarily used to transform data which violated normality assumption so that the transformed data can conforms normality assumption. Therefore there are two kind of normality assumption used in statistics, those are univariate normal and multivariate normal. Maximum likelihood method for λ estimation in Box-Cox transformation distinguishes the estimation formula for univariate and multivariate normal. Box-Cox transformation for multivariate normal is more complicated than the univariate formula. In practice sometime Box-Cox formula for univariate normal is used to get estimation of λ for multivariate normal because of the complexity in calculation. In this paper we will cover λ estimation for both univariate and multivariate case.

2.1.1. Univariate

For univariate case, the appropriate power λ is the solution for maximize the expression

$$l(\lambda)=-\frac{n}{2}\ln\left|\frac{1}{n}\sum_{i}\left(x_i^{(\lambda)}-\overline{x^{(\lambda)}}\right)^2\right|+(\lambda-1)\sum_{i}\ln x_i \qquad (4)$$

Where

$$\overline{x^{(\lambda)}}=\frac{1}{n}\sum_{i}x_i^{(\lambda)}=\frac{1}{n}\sum_{i}\left(\frac{x_j^{\lambda}-1}{\lambda}\right) \qquad (5)$$

$x^{(\lambda)}$ is the arithmetic average of the transformed observation.

2.1.2. Multivariate

In multivariate observation, power of transformation can be searched for each variable with equation. Let $\lambda_1,\lambda_2,\ldots,\lambda_p$ are power transformation from p variable in data. Then each λ_k are found by maximize

$$l_k(\lambda)=-\frac{n}{2}\ln\left|\frac{1}{n}\sum_{j}\left(x_{jk}^{(\lambda_k)}-\overline{x^{(\lambda_k)}}\right)^2\right|+(\lambda_k-1)\sum_{j}\ln x_{jk} \qquad (6)$$

Where x_{1k}, x_{2k}, …, x_{nk} are n*th* observation from k*th* variable.

$$\overline{x^{(\lambda_k)}}=\frac{1}{n}\sum_{j}x_{jk}^{(\lambda_k)}=\frac{1}{n}\sum_{j}\left(\frac{x_{jk}^{\lambda_k}-1}{\lambda_k}\right) \qquad (7)$$

$\mathbf{x}^{(\lambda_k)}$ is the arithmetic average of the transformed observation.

2.2. Nelder Mead Algorithm

According to J.A. Nelder and R. Mead (1965), Nelder Mead Simplex Direct Search is defined as "A method is described for minimization of function of n variables, which depends on the comparison of function values at the (n+1) vertices of a general simplex, followed by the replacement of the vertex with highest value by another point. The simplex adapts itself to the local landscape, and contract on to the final minimum. The method is shown to be effective and computationally compact. A procedure is given for the estimation of the Hessian Matrix in the neighborhood of the minimum, needed in statistical estimation problems."

Nelder Mead Algorithm require us to define the value of parameters such as coefficients of reflection (ρ), expansion (χ), contraction (γ), and shrinkage (σ) before minimizing a function. The value of those parameters must satisfy

$$\rho>0,\ \chi>1,\ \chi>\rho,\ 0<\gamma<1,\ and\quad 0<\sigma<1$$

The value of coefficients used in Nelder Mead are flexible as long as the parameters conditions are satisfied. This could lead into minor difference between Nelder Mead algorithm used by different author. In this paper we use common parameters value for the algorithm where

$$\rho=1,\ \chi=2,\ \chi>\rho,\quad \gamma=\frac{1}{2},\ and\quad \sigma=\frac{1}{2}$$

According to Lagarias et al. the Nelder Mead algorithm steps for one iteration are

1. **Order**. Order the n+1 vertices to satisfy $f(x_1)\leq f(x_2)\leq\ldots\leq f(x_{n+1})$
2. **Reflect**. Compute the reflection point x_r from

   $$x_r=\bar{x}+\rho(\bar{x}-x_{n+1})=(1+\rho)\bar{x}-\rho x_{n+1}$$

   Where $\bar{x}=\sum_{i=1}^{n} x_i/n$ is centroid from n best point (all vertices except for x_{n+1}). Evaluate $f_r=f(x_r)$. If $f_1\leq f_r<f_n$, accept the reflected point x_r and terminate the iteration.
3. **Expand**. If $f_r<f_1$, calculate the expansion point x_e.

   $$x_e=\bar{x}+\chi(x_r-\bar{x})=\bar{x}+\rho\chi(\bar{x}-x_{n+1})=(1+\rho\chi)\bar{x}-\rho\chi x_{n+1}$$

   And evaluate $f_e=f(x_e)$. if $f_e<f_r$, Accept x_e and terminate the iteration; otherwise (if $f_e\geq f_r$) accept x_r dan terminate the iteration.
4. **Contract**. if $f_r<f_n$ perform a contraction between $\bar{x}$ and the better of x_{n+1} and x_r

   a. **Outside**. If $f_n\leq f_r<f_{n+1}$, perform an outside contr*action* : calculate

   $$x_c=\bar{x}+\gamma(x_r-\bar{x})=\bar{x}+\gamma\rho(\bar{x}-x_{n+1})=(1+\rho\gamma)\bar{x}+\rho\gamma x_{n+1}$$

   And evaluate $f_c=f(x_c)$. If $f_c\leq f_r$ accept x_c stop iterasi. Otherwise, go to step 5 (*shrink step*).

   b. **Inside**. If $f_r\geq f_{n+1}$, perform inside contr*action* : calculate

   $$x_{cc}=\bar{x}+\gamma(\bar{x}-x_{n+1})=(1-\gamma)\bar{x}+\gamma x_{n+1}$$

   c. And evaluate $f_{cc}=f(x_{cc})$. If $f_{cc}<f_{n+1}$ accept x_{cc} and terminate the iteration; Otherwise, go to step 5 (*shrink step*).

5. **Perform a shrink step**. Evaluate f at the n point $v_i = x_1 + \sigma(x_i - x_1)$, i=2,…,n+1. The unordered vertices of simplex at the next iteration consist of $x_1, v_2, \dots, v_{n+1}$

2.3. Quasi Newton BFGS

Quasi newton is continued development of newton method that estimate hessian matrix instead of using exact differentiation. Hessian matrix used when function that needed to get optimum value consist of more than one variable which denote by :

$$H(f)= \left| \frac{\partial^2 f}{\partial x_2 \partial x_1} \quad \frac{\partial^2 f}{\partial x_2^2} \quad \cdots \quad \frac{\partial^2 f}{\partial x_2 \partial x_n} \right|$$

There is advantage of estimate hessian matrix. Optimum value for unfixed number variables function will be easier to get because there’s no need to search differentiate function for different variable explicitly just to get the hessian matrix. The classical newton method use following equation to search minimum value of function

$x_{i+1}=x_i-H_i^{-1}\nabla f$

Many approximations can be used to estimate hessian matrix, one of them that frequently used is BFGS method. So, Quasi newton BFGS basically is method to get optimum value of function where to estimate hessian matrix using BFGS method. Kurt, Bryan, and Zak bring up the algorithm of finding optimum value using quasi newton method:

1. Initiate initial value x_0 as initial predicted minimum point; i=0; $H_0 = I$(initial hessian matrix is identity matrix with pxp size, where p= number of variables).
2. Count $\nabla f(\mathbf{x}_i)$ and search direction $h_i = -H_i \nabla f(x_i)$.
3. Using Line search to search value of $x_{i+1} = x_i + t^* h_i$ where t^* is a value that minimize $f(x_i + t^* h_i)$
4. Count H_{i+1} where $H_{i+1} = H_i + U_i$ and U_i is the update of hessian matrix. In this paper, we use BFGS method to get U_i (Note there are several different updating rule for Hessian such as DFP etc).

Hessian matrix using BFGS method update by

$$H_{i+1} = H_k - \frac{\delta_i \gamma_i^t H_i + H_i \gamma_i \delta_i^t}{\delta_i^T \gamma_i} + \left(1 + \frac{\gamma_i^t H_i \gamma_i}{\delta_i^T \gamma_i}\right) \frac{\delta_i \delta_i^t}{\delta_i^T \gamma_i}$$

where $\delta_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i,\ \gamma_i = \boldsymbol{g}_{i+1} - \boldsymbol{g}_i,\ \boldsymbol{g}_i = \nabla f(\boldsymbol{x}_i)$

3. Simulation Result

We use three dataset in this simulation to provide an overview how well the two algorithm solve the Box-Cox transformation (both univariate and multivariate case). First Simulation use radiation data of door closed and door open from “Applied Multivariate Statistics 5th Edition”, page 181. Second simulation use generated data with 500 sample size. The generated data were taken from gamma distribution ($\alpha=7, \beta=5$), exponential distribution ($\beta=2$), and squared normal distribution ($N(\mu = 10 , \sigma = 50)^2$). In third simulation we use survey data from BPS (SUSENAS kor triwulan 3 2011 blok 43). The third simulation use food expenditure, non-food expenditure, income per capita, and expenditure as variable. All of variables in third simulation are aggregative data sort by province in Indonesia.

In this simulation we use Java and R to generate data and running the simulation process. Both of optimization algorithm using 0 as initial minimum value.

3.1. Nelder Mead vs Quasi Newton BFGS Performance

3.1.1. Radiation Data

Table 1 Comparison of Function Evaluation between Quasi Newton BFGS and Nelder Mead

Variable	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	Nelder mead	BFGS	Nelder mead
Door Closed	15	30	16	35
Door Open	21	32		

Table 2 Comparison of Elapsed Time (in second) between Quasi Newton BFGS and Nelder Mead

Simulation	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	*Nelder mead*	BFGS	*Nelder mead*
1	0.019331485	0.028886779	0.034593148	0.059916183
2	0.038517302	0.028886779	0.026931542	0.06467263
3	0.016545322	0.036451957	0.032389658	0.062368608
4	0.020559749	0.029267022	0.027601753	0.063092847
5	0.016247146	0.020382622	0.039609473	0.027093623
6	0.014794566	0.01534031	0.027488227	0.063090795
7	0.015282863	0.029325152	0.028337616	0.063064806
8	0.019220011	0.035298237	0.034227267	0.058602432
9	0.014209158	0.014334993	0.027136024	0.031490344
10	0.015124201	0.029288222	0.034184866	0.063712449
Average	0.01898318	0.026746207	0.03125	0.05571

3.1.2. Generated Data

Table 3 Comparison of Function Evaluation between Quasi Newton BFGS and Nelder Mead

Variable	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	Nelder mead	BFGS	Nelder mead
V1 (gamma)	11	36	31	82
V2 (exponential)	23	36		
V3 (chi square)	15	38		

Table 4 Comparison of Elapsed Time (in second) between Quasi Newton BFGS and Nelder Mead

Simulation	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	*Nelder mead*	BFGS	*Nelder mead*
1	0.055284888	0.115203806	0.229305866	0.455354371
2	0.066271561	0.119212762	0.25443536	0.470266567
3	0.056126071	0.107775405	0.244678319	0.463293636
4	0.061207365	0.118074087	0.234084881	0.467130936
5	0.056615736	0.100893432	0.240787676	0.483166761
6	0.055364219	0.118448174	0.223415532	0.459529512
7	0.059190577	0.11106833	0.242823612	0.477459709
8	0.054575695	0.099898374	0.257206478	0.457195399
9	0.057264746	0.108566665	0.237874993	0.478134023
10	0.062617543	0.111782994	0.259513919	0.476619209
Average	0.05845184	0.111092403	0.242412664	0.468815012

3.1.3. SUSENAS Data

Table 5 Comparison of Function Evaluation between Quasi Newton BFGS and Nelder Mead

Variable	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	Nelder mead	BFGS	Nelder mead
V1	9	30	432	232
V2	21	30		
V3	9	32		
V4	18	30		
V5	14	32		

Table 6 Comparison of Elapsed Time (in second) between Quasi Newton BFGS and Nelder Mead

Simulation	Marginal (Univariate Box-Cox)		Simultaneous (Multivariate Box-Cox)	
	BFGS	*Nelder mead*	BFGS	*Nelder mead*
1	0.027268698	0.066015103	0.589151061	0.435189909
2	0.025188309	0.064844969	0.602391832	0.435628966
3	0.024275317	0.064827872	0.590433353	0.433271601
4	0.03300037	0.071862353	0.587689591	0.440749925
5	0.026002137	0.066482199	0.598696098	0.438091649
6	0.025440663	0.064880531	0.589017019	0.436830559
7	0.030915877	0.066659326	0.603449809	0.433643637
8	0.028483969	0.071748826	0.594827339	0.445973468
9	0.025499478	0.064926352	0.595516014	0.440810107
10	0.028265125	0.067280981	0.592387223	0.441822263
Average	0.027433994	0.066952851	0.594355934	0.438201208

4. Conclusion

In this paper we have focused to comparing efficiency of quasi-newton BFGS and Nelder Mead algorithm applied to Box-cox transformation. Generally quasi newton BFGS perform more efficient than Nelder Mead algorithm in terms of number of functions evaluations and number of iteration. However, the time differences among those two algorithm is slightly different. Simulation using SUSENAS's data result in condition where Nelder Mead algorithm is more efficient than quasi-newton BFGS in case of multivariate data transformation. From the simulation, we can conclude that generally quasi newton BFGS perform more efficient than Nelder Mead when applied in Box-Cox transformation formula. However the time elapsed and evaluation function difference between the two algorithm sometime are so small and become negligible. Therefore, both algorithm can be used side by side to overcome the weakness of each method, then case which failed to convergent can be minimalized.

In this research we also did normality test before and after the Box-Cox transformation applied to dataset. For testing univariate normal assumption, we use liliefors and Shapiro wilks test. Meanwhile for testing multivariate normal assumption, we use multivariate shapiro wilks and mardia kurtosis test. All dataset in this paper is not conform both univariate and multivariate normal assumption before Box-Cox transformation applied. In Radiation Data after univariate (marginal) Box-Cox transformation applied to it, liliefors test result in accept H_0 for *Door Closed Variable* (passed univariate normal test) and reject H_0 for *Door Open Variable* (didn't pass univariate normal test), however when Shapiro wilks test used against transformed data all variable are passed univariate normal test. Furthermore when multivariate (simultaneous) Box-Cox transformation applied to Radiation Data, both shapiro wilks and mardia kurtosis test result in accept H_0 thus transformed data is multivariate normal according to the test. For the rest of dataset (Generated Data, Susenas Data), Box-Cox transformation make data passed univariate and multivariate normal assumption except for V3 (chi square) variable from generated data which failed to conform the univariate normal assumption and Susenas Data which failed to conform the multivariate normal assumption. According to the simulation in this research the Box-Cox transformation can remedy non-univariate and non-multivariate normal data to conform univariate normal and multivariate normal distribution well enough. As a note the different univariate normal and multivariate normal test can be produce a different conclusion in the end. According to the simulation the Box-Cox transformation also is not guaranteed the transformed data to pass the normality test, the more data resembled normal distribution before transformation (i.e. skewness of data, kurtosis of data), the more transformed data to conform the univariate normal and multivariate normal distribution.

References

[1] Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics* (Vol. 4). Belmont, CA: Duxbury Press.

[2] Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistikal Society. Series B (Methodological),* 211-252.

[3] Chapra, S. C., & Canale, R. P. (2010). *Numerical methods for engineers*. *New* York: McGraw-Hill Higher Education.

[4] Cleveland, W. S. (1984). Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging. *The American Statistician*, 38(4), 270-280.

[5] Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder--Mead simplex method in low dimensions. *SIAM Journal on optimization*, *9*(1), 112-147.

[6] Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, *7*(4), 308-313.

[7] Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. *Practical Assessment, Research & Evaluation*, *15*(12), 1-9.

[8] Sakia, R. M. (1992). The Box-Cox transformation technique: a review. *The statistikian*, 169-178.

[9] Tukey, J. W. (1957). On the comparative anatomy of transformations. *The Annals of Mathematical Statistics*, 602-632.

[10] Chong, E. K., & Zak, S. H. (2013). *An introduction to optimization* (Vol. 76). John Wiley & Sons.