

A feasibility study in adapting Shamos Bickel and Hodges Lehman estimator into T-Method for normalization

N Harudin^{1,2}, K R Jamaludin¹, M Nabil Muhtazaruddin¹, F Ramlie¹, Wan Zuki Azman Wan Muhamad¹

¹ Razak School in Engineering & Advanced Technology Universiti Teknologi Malaysia, 54100 Kuala Lumpur, Malaysia

² Department of Mechanical Engineering, Universiti Tenaga Nasional, 43000 Kajang Selangor, Malaysia

nolia.harudin@gmail.com, bkhairur.kl@utm.my

Abstract. T-Method is one of the techniques governed under Mahalanobis Taguchi System that developed specifically for multivariate data predictions. Prediction using T-Method is always possible even with very limited sample size. The user of T-Method required to clearly understanding the population data trend since this method is not considering the effect of outliers within it. Outliers may cause apparent non-normality and the entire classical methods breakdown. There exist robust parameter estimate that provide satisfactory results when the data contain outliers, as well as when the data are free of them. The robust parameter estimates of location and scale measure called Shamos Bickel (SB) and Hodges Lehman (HL) which are used as a comparable method to calculate the mean and standard deviation of classical statistic is part of it. Embedding these into T-Method normalize stage feasibly help in enhancing the accuracy of the T-Method as well as analysing the robustness of T-method itself. However, the result of higher sample size case study shows that T-method is having lowest average error percentages (3.09%) on data with extreme outliers. HL and SB is having lowest error percentages (4.67%) for data without extreme outliers with minimum error differences compared to T-Method. The error percentages prediction trend is vice versa for lower sample size case study. The result shows that with minimum sample size, which outliers always be at low risk, T-Method is much better on that, while higher sample size with extreme outliers, T-Method as well show better prediction compared to others. For the case studies conducted in this research, it shows that normalization of T-Method is showing satisfactory results and it is not feasible to adapt HL and SB or normal mean and standard deviation into it since it's only provide minimum effect of percentages errors. Normalization using T-method is still considered having lower risk towards outlier's effect.

1. Introduction

The theory of Mahalanobis Taguchi System (MTS) was inspired from a statistical tool called Mahalanobis Distance (MD) which introduced by a famous Indian statistician named Dr. Prasanta Chandra Mahalanobis in 1936. T-Method is one of various tool govern under MTS theory which specifically developed to calculate an overall prediction based on the signal to noise ratio to predict the future value based on old and current data together. Unique characteristic of T-method is the application of signal to noise ratio as measurement for the function robustness of target system. MTS including T-Method considered to be an ad hoc in the sense that they are not developed using any



underlying statistical and probability theory as well as not been compared to any other multivariate statistical method. MTS is not based on probability and distribution theory since the theory developed only measure the descriptive statistic as agreed by [1]–[3]. The normalization stage in T-method for instance, is relying on the differences between unit space and signal data only without considering any statistical theory even the mean and standard deviation. The advantages of T-Method are its simplicity to be understood as well as its ability to do a prediction with a very limited sample size. It is very important for the user of T-Method to clearly understand the data since the method is not considering the effect of outliers within the population data.

In most cases, assumption that receives much attention from most of the statistician is that the regression analysis must be free from the effect of outliers. The errors and distribution was assumed to be normally distributed, observations are random, independent and identically distributed and equally reliable with no outlier in the data. Hodges-Lehmann estimator (HL) is defined as the median of the pairwise Walsh averages and been introduced since 1963 in various application such [4]–[11]. The main advantage of the HL estimator is that it is robust against outliers in a sample. It has a breakdown point of 0.29 (29%) is the least portion of data contamination needed to derive the estimate beyond all bounds). If the underlying distribution for the data is normal, then the asymptotic relative efficiency (ARE) of the HL estimator relative to the sample mean is 0.955 otherwise, it is often greater than unity[12]. It is well known that the classical standard deviation is not robust to outliers, where even a single outlier can have a severe effect[13]. The Shamos Bickel (SB) is an analogous scale estimator to the location estimator of Hodges-Lehmann. The square of the SB estimator can be used as an estimate for variances. Of course, SB is not unbiased or median unbiased for standard deviation, SB^2 is not unbiased or median unbiased for σ^2 . But they give good approximations to what they are estimating [14]. HL and SB are comparable method to normal mean and standard deviation which been applied in various area with intention of considering the outliers effect. The application of HL and SB is very well understood in control chart improvement as discussed by many researchers [12], [14]–[16].

This study was conducted to see the feasibility of embedding robust statistical estimators into T-Method for enhancing the accuracy of normalization steps especially towards outliers which currently T-Method is not relying on any statistical inference analysis. This study is limited to 2 different case studies with different sample size and number of unknown data for prediction. The result is compared based on the prediction of unknown data towards the actual value. The range and mean of the error percentages will be the comparison parameters as well as the R-Squared value. The conclusion driven from this study is mainly applied for these 2 case studies only since more examples are needed for firm fact in determining the robustness of T-Method.

2. Methodology

The T-Method concept is briefly explained in this chapter as well as the Hodges-Lehmann estimator and Shamos Bickel estimator. Most of the literatures are normally pronounced both estimator as Hodges-Lehmann location estimator and Shamos Bickel scale estimator which make them comparable to the theory of mean and standard deviation. The study had been conducted and analysed using two different case studies on prediction output.

2.1. T-Method

T-Method calculation procedure is a combination of Mahalanobis Distance concept with S/N Ratio (SNR). The theory behind T-Method is still focusing on three important elements of SNR which are sensitivity, linearity and variability. Considering on the dynamic environment adding some value added towards the theory developed. T-Method 1 as well applying the concept of reference-point proportional equation which created a linear regression that passes through origin of the graph generated. Unit group and signal group are used to establish and validate forecasting mode. Average value of unit group are subtract from each member of signal group for the normalization.

To find out which of the items (parameter) will be useful for prediction and estimation, an item by item computation of the proportional coefficient (β) and SN ratio (η) is performed. In T-Method, proportional coefficient (β) and SN ratio (η) are calculated between the output value (M_i) and the item value. The computation of the proportional coefficient (β) and SN ratio (η) is performed item by item

with the use of normalized data (X_{ij}) and normalized output value (M_i). The SN ratio η and proportional coefficient β are found for each item respectively. If the SN ratio η turns out to be negative, the value will be treated as zero. The integrated estimate output value for Signal Data is found using the proportional coefficient β and SN ratio η , item by item. The higher the SN ratio of an item, the greater the degree of its contribution to general output estimation. Integrated estimate output value \hat{M} can be found using the following formula:

$$\hat{M}_i = \frac{\eta_1 \times \left(\frac{X_{i1}}{\beta_1}\right) + \eta_2 \times \left(\frac{X_{i2}}{\beta_2}\right) + \dots + \eta_j \times \left(\frac{X_{ij}}{\beta_j}\right)}{\eta_1 + \eta_2 + \dots + \eta_j} \quad (2.1)$$

Formulas to calculate the SN ratio η (db) for integrated estimate values are shown below which is used to calculate the degree of contribution:

$$\text{Linear equation, } L = M_1 \hat{M}_1 + M_2 \hat{M}_2 + \dots + M_i \hat{M}_i \quad (2.2)$$

$$\text{Effective Divider, } r = M_1^2 + M_2^2 + \dots + M_i^2 \quad (2.3)$$

$$\text{Total Variation, } S_T = \hat{M}_1^2 + \hat{M}_2^2 + \dots + \hat{M}_i^2 \quad (2.4)$$

$$\text{Variation of proportional term, } S_\beta = \frac{L^2}{r} \quad (2.5)$$

$$\text{Error variation, } S_e = S_T - S_\beta \quad (2.6)$$

$$\text{Error Variance, } V_e = \frac{S_e}{n-1} \quad (2.7)$$

$$\text{SN ratio, } \eta = 10 \log \left(\frac{(S_\beta - V_e)}{r V_e} \right) \quad (2.8)$$

Through the computations performed so far, output values have been normalized in terms of the average Unit Space value. Therefore, in order to find the integrated estimate value (\hat{Y}) of the actual output, the average value (M_0) of the output of the Unit Space is added to the normalized value (\hat{M}). For instance, the integrated estimate value (\hat{Y}_1) for the output of Signal Data No.1 is found as follow (similarly for unknown data):

$$\hat{Y}_1 = \hat{M}_1 + M_0 \quad (2.9)$$

2.2. Hodges Lehman estimator and Shamos Bickel estimator

The Hodges-Lehmann (HL) estimator was proposed by Hodges and Lehmann in 1967 as an estimator for the point of symmetry θ of a continuous and symmetric distribution. It is used for the estimation of the location parameter in one-sample and two-sample models[15]. In this study, only one-sample model is used since the intention is to compare HL to the mean of the distribution data.

For that case, let us assume a random variables of X_j ($1 \leq j \leq m$) and Hodges-Lehmann estimator (HL) of the expected value $E(X) = E(X_i)$ (the one-sample problem) is follow:-

$$E^{HL}(X) = \text{median} \left(\frac{x_j + x_k}{2} \right) \quad (2.10)$$

where $1 \leq k \leq m$ [4]

To make the standard deviation (SD) comparable to Shamos Bickel scale estimator (SB), we consider the normalized SBSE as

$$SB(x) = \text{Median} \left(\frac{|x_j - x_k|}{0.9538726} \right) \quad (2.11)$$

Both the HL and SB will be used as a replacement of unit space and signal data deduction steps in calculating the normalization. The normal mean and standard deviation also been applied as part of comparison purposes.

2.3. Data Collection, equipment and tool setup

Data on power consumption prediction is taken from a real case study conducted at nuclear agency Malaysia which involved 15 variables and 14 sample size. The second data is taken from the [17] reference case study on yield% prediction which involving very minimum sample data (6 variables and 7 sample size). In order to test the robustness of each method, extreme outliers' data are purposely added to the population data to see the impact of each method compared. The prediction of unknown data for power consumption involved 13 sample data while yield% prediction is only for 1 data.

As for a comparable study, the formulation of T-Method including the effect of HL, SB, normal mean and standard deviation had been constructed using Matlab R2013A application software. HL and SB is calculated using R programming to get the output prior adapting the result into Matlab.

3. Result and Analysis

In this feasibility study, the main idea is to see how robust T-Method is when it comes to an existence of outliers in the data population. It is believe that by knowing the data trend very well, the effect of outliers might not exist at all since a crucial decision might done earlier. However a deep understanding on the data series is required for that purpose which might be a risk for certain people and can be considered not to be robust enough. As stated earlier, the analysis is involving 2 difference set of case study. Each of the case studies is then differentiate into 2 difference case of error% analysis which is with and without extreme outliers' consideration. The actual data of the unknown trend is taken for a comparison purposes.

Table 1 shows that without adding the extreme outliers into the data, the accuracy shown by HL and SB is much better compared to T-Method, Normal mean and standard deviation which on average error percentages of 4.67%. However, when extreme outliers were added, T-Method is showing higher accuracy compared to others with average error percentages of 3.09%. The R squared value for all method is not vary much and still showing high correlation between predicted and actual data. If compare the range of the error percentages for each method in table 1, T-method is showing minimum range compared to others. This highlighted that the prediction accuracy is having lower gap among samples.

The second case study on yield percentages prediction which summarize in table 2, are showing that HL and SB is having minimum error with extreme outliers case while T-Method is the minimum for case without extreme outliers. This is completely different compared to prediction of power consumption summarize in table 1. There is no range percentages analysis on this since the prediction is involving only 1 unknown data. A few things need to be pondered and question further is the total number of sample size considered in both cases. T method claim to be a good prediction tool for a minimum sample size data which expect no extreme outliers in it and it proved from the yield% prediction case. Minimum sample size can easily trace the dispersion of the data, so normality of it is easily controlled as well as the outliers. By adding extreme outliers on it purposely might be irrelevant to normal practise, however if its involved huge sample size, outliers might be not easily been controlled and normality of the data as well is hardly defined unless the data is really normal or else adjustment and assumption on normality is the usual practice. As for this study, it shows that T-Method is more accurate in doing prediction in case of minimum sample size as well as moderate sample size compared to Hodges Lehman and Shamos Bickel. Since the result is relying mainly on this study, future analysis on data with higher sample size is needed for better comparison.

Table 1. Prediction Vs. Actual data of power consumption

Normalization Method	Error% range for predicted and actual data (without extreme outliers)	R-Squared predicted Vs actual (without extreme outliers)	Error% range for predicted and actual data (with extreme outliers)	R-Squared predicted Vs actual (with extreme outliers)
T-Method (unit space)	Range:0.98%~8.85% Average: 6.50%	75.65%	Range:0.67%~7.29% Average: 3.09%	71.61%
HL and SB estimator	Range: 0.35% ~ 9.75% Average: 4.67%	77.67%	Range:0.13%~9.67% Average: 4.48%	69.85%
Normal mean & SD	Range:0.52% ~ 10.83% Average: 5.42%	77.27%	Range: 0.63%~11.19% Average: 5.48%	72.17%

Table 2. Prediction Vs. Actual data of yield % prediction

	Normalization Method	Estimated	Actual	Error%
Yield% prediction (Without extreme outliers)	T-Method	75.13		2.50%
	normal mean & SD	82.4725	73.3	12.51%
	HL & SB	82.3043		12.28%
Yield% prediction (With extreme outliers)	T-Method	83.7827		14.30%
	normal mean & SD	80.4	73.3	9.69%
	HL & SB	80.0028		9.14%

4. Conclusion

Comparing results in both Table 1 and Table 2 and by considering the irrelevant of adding the extreme outliers on the minimum sample size in yield percentage prediction case, it shows that T-Method is perform better than other 2 methods. However, relying on these 2 case studies might not be sufficient enough to summarize the overall effectiveness of robust normalization in T-method. This result is proved to be effective mainly on these 2 case studies only. Further analysis should be done to several other prediction cases so that the robustness of normalization in T-Method can be summarized in more accurate manner. As for this study, T-method was summarized to be more accurate in doing prediction compared to HL & SB and Normal mean & SD. Therefore, it is not feasible to adapt the Shamos Bickel and Hodges Lehman into T-method for the normalization analysis within this study. However it is very useful to have HL and SB as a comparable method dealing with normalization process. Normalization using T-method is still considered having lower risk towards outlier's effect.

Acknowledgments

Authors wish to acknowledge Mr. Yusnan from nuclear agency Malaysia for his willingness in supporting this research by providing guidance and permission in data collection stage. Special thanks to ministry of higher education (MoHE, Malaysia) and Universiti Tenaga Nasional for their financial support throughout the author research duration period. This research paper publication is mainly supported by Tier 1 Research University Grant No: 15H43 (Universiti Teknologi Malaysia).

References

- [1] K. Tsui, T. Sukchotrat, and V. C. P. Chen, "A comparison study and discussion of the Mahalanobis-Taguchi System Seoung Bum Kim," *Int. J. Ind. Syst. Eng.*, vol. 4, no. 6, pp. 631–644, 2009.
- [2] W. H. Woodall, R. Koudelik, K. L. Tsui, S. B. Kim, Z. G. Stoumbos, and C. R. Carvounis, "A review and analysis of the Mahalanobis-Taguchi system," *Technometrics*, vol. 45, no. 1, pp. 1–15, 2003.

- [3] S. Taguchi, R. Jugulum, G. Taguchi, and J. O. Wilkins, "Discussion," *Technometrics*, vol. 45, no. 1, pp. 16–21, 2003.
- [4] R. Duchnowski and Z. Wiśniewski, "Accuracy of the Hodges–Lehmann estimates computed by applying Monte Carlo simulations," *Acta Geod. Geophys.*, 2016.
- [5] C. H. Park, S. Lee, and J. H. Chang, "Robust closed-form time-of-arrival source localization based on α -trimmed mean and Hodges-Lehmann estimator under NLOS environments," *Signal Processing*, vol. 111, pp. 113–123, 2015.
- [6] I. A. Canay and T. Otsu, "Hodges-Lehmann optimality for testing moment conditions," *J. Econom.*, vol. 171, no. 1, pp. 45–53, 2012.
- [7] E. Inference and D. Adminl, "Inaccuracy rates and Hodges-Lehmann large deviation rates for parametric inferences with nuisance parameters," vol. 8, 1995.
- [8] Y. Y. Nikitin, "Hodges-Lehmann and Chernoff efficiencies of linear rank statistics," *J. Stat. Plan. Inference*, vol. 29, no. 3, pp. 309–323, 1991.
- [9] J.N.Adichie, "Estimates of Regression Parameters Based on Rank Test," *Ann. Math. Stat.*, vol. 38, no. 3, pp. 894–904, 1967.
- [10] Hodges Jr, J.L. and Lehmann, E.L., "Estimates of location based on rank tests," *Ann. Math. Stat.*, p. pp.598-611, 1963.
- [11] P. j. B. E.L.Lehmann, "Descriptive statistics for nonparametric models. III. Dispersion," *Ann. Stat. @ www.jstor.org*, vol. 4, no. 6, pp. 1139–1158, 1976.
- [12] H. Z. Nazir, M. Riaz, R. J. M. M. Does, and N. Abbas, "Robust CUSUM control charting," *Qual. Eng.*, vol. 25, no. 3, pp. 37–41, 2013.
- [13] G. Tarr, N. C. Weber, and S. Müller, "The difference of symmetric quantiles under long range dependence," *Stat. Probab. Lett.*, vol. 98, pp. 144–150, 2015.
- [14] M. O. Abu-shawiesh, M. B. Abdullah, Q. Methods, and S. Arabia, "A New Robust Bivariate Control Chart for Location," *Commun. Stat. Comput.*, no. March 2012, pp. 37–41, 2007.
- [15] M. O. Abu-Shawiesh and M. B. Abdullah, "New Robust Statistical Process Control Chart for Location," *Qual. Eng.*, vol. 12, no. 2, pp. 149–159, 1999.
- [16] A. A. Aly and Aydin ozturk, "Hodges-Lehmann quantile-quantile plots," *Comput. Stat. Data Anal.*, vol. 6, pp. 99–108, 1988.
- [17] S. Teshima, Y. Hasegawa, and K. Tatebayashi, *Quality Recognition and Prediction: Smarter Pattern Technology with the Mahalanobis-Taguchi System*. 2012.