The Analysis of a Fractional Factorial Experiment With Missing Data Using Neural Networks

Taho Yang
Chao-Ton Su

Abstract

Fractional factorial design has been widely used to generate a robust design in many industrial settings. This paper presents a neural network approach for missing data estimation for the analysis of a fractional factorial experiment. Empirical results illustrate that the proposed methodology is both effective and efficient in providing a quality data estimation for a variety of scenarios towards a fractional factorial experimental design. The proposed methodology could be an alternative to commonly used statistical approaches in that it features both the easy implementation and the learning capability of a neural net.

Keywords: Factorial experiment, Taguchi method, prediction, missing data, neural network.

1 INTRODUCTION

A *factorial experiment* has been used in many industrial applications to generate a *robust design* which is an engineering methodology for improving productivity during research and development (R&D), so that high-quality products can be produced quickly and at low cost (Phadke 1988). When a *full factorial experimental design* is not achievable due to costly data collection process and/or physical limitations, a *fractional factorial experiment* becomes much more practical in achieving design objectives. Taguchi's experimental design method marks the milestone in this research area (Box 1988, Kachar 1985, Ross 1996, Taguchi *et al.* 1989). While the fractional factorial design is remarkable in generating a robust design for an industrial application, it fails to complete an experiment when a missing-data situation occurs. Missing data situation is often inevitable due to a variety of reasons such as: lost data, cost limitations, experimental interruptions, *etc.*, therefore, a significant amount of researches have been aiming at reducing the influences of the missing data situation through statistical inferences (Enders 2010, Enders and Bandalos 2001, Graham 2009, Little and Rubin 2014).

Statistical methodologies, such as *linear regression* (*LR*) and *maximum likelihood estimator* (*MLE*) approaches, have been being developed in literature for missing data analysis in the past few decades. The main advantage of a statistical method is that it yields an unambiguous model of the estimation process with characteristics that can be evaluated analytically (DeSarbo and Green 1986). However, this analytical power is bought at the price of rather stringent distributional assumptions about the population (Gleason and Staelin 1975). In addition, they often require sophisticated statistical backgrounds for the data analysis process. For other estimation methods, *e.g.*, average value, data elimination, *etc.*, which are easier to implement but is lack of proven solution quality.

This paper proposes a neural network approach to solve a missing data factorial experiment. Neural network is a proven tool in searching a quality solution through learning (Dayhoff 1990, Fausett 1994, Stern 1996). The neural net model estimates the missing data values through learning. The estimated values are then used to complete factorial experiment design. Its solution quality is justified through benchmarking against existing commonly used approaches.

The remainder of this paper is organized as follows. The proposed solution procedure is discussed in section 2. In section 3, numerical results are provided to illustrate the procedure and to demonstrate its efficiency and effectiveness. It is then followed by conclusions in section 4.

2 SOLUTION METHODOLOGY

The proposed solution methodology uses the neural net to estimate the missing data that in combination with existing data are then used to conduct fractional factorial experiment to find the optimal control factor levels. The overall procedure is shown as Figure 1.

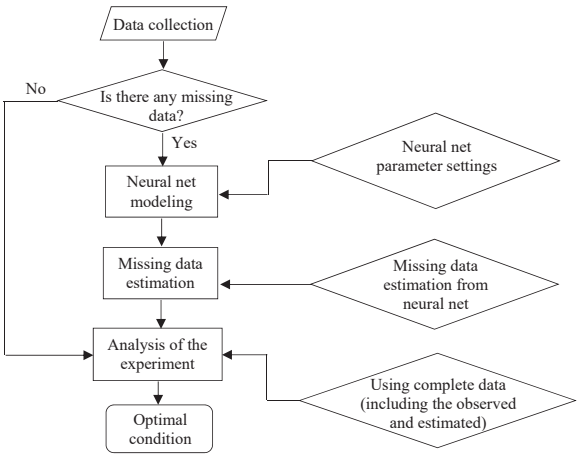

Figure 1. Proposed methodology

The proposed methodology begins with the data collection process. When there is not any missing data, it proceeds with the analysis of the experiment. Otherwise, it calls for neural net modeling for response value prediction. Backpropagation (BP) neural network is constructed. For a detailed discussion leading towards the choice of the BP network for our application, readers are referred to [13-16]. Key parameters including momentum, learning rate, number of hidden layers, and stopping criterion are determined through experiments.

Following the completion of the neural net modeling, the missing data are predicted through the trained neural net. At the end of this step, the complete data set is available for the experiment analysis. The optimal control factor level combination can then be determined.

3 COMPUTATIONAL RESULTS

A numerical example of Taguchi's orthogonal array L_8 (2^5) is first adapted from Su and Miao (1998) to illustrate the proposed procedure's effectiveness. This design aims to minimize the response value. Table 1 contains the original data set of the experiment. Each trial has two response values from two repetitions. Through Taguchi's experiment design method (Phadke 1988), this data set leads to an optimal combination of $A_2C_2D_2E_1$. Note that factor B is not significant.

Table 1. Illustrative data set

No.	Control factors					Responses	
	A	B	C	D	E		
1	1	1	1	1	1	66	66
2	1	1	2	2	2	68	63
3	1	2	1	2	2	80	88
4	1	2	2	1	1	63	65
5	2	1	1	1	2	73	71
6	2	1	2	2	1	37	42
7	2	2	1	2	1	38	39
8	2	2	2	1	2	57	48

From the above example, we randomly delete 15% (3 data points) out of the original 16 data to create a missing data example, case 1, as shown in Table 2. Table 2 is analyzed by our proposed procedure.

Table 2. Data for case 1 with 5% missing data

No.	Control factors					Responses	
	A	B	C	D	E		
1	1	1	1	1	1	66	66
2	1	1	2	2	2	–	63
3	1	2	1	2	2	–	88
4	1	2	2	1	1	63	65
5	2	1	1	1	2	73	–
6	2	1	2	2	1	37	42
7	2	2	1	2	1	38	39
8	2	2	2	1	2	57	48

Step 1 of the proposed procedure detects the presence of the missing data. Next step proceeds with the neural net modeling. Among the available data points, 10 of 13 are randomly chosen for neural net training samples and the remainders of them are used for testing samples. A commercial neural net modeling software, Qnet97 (Qnet97 User's manual 1997), is used to aid the modeling process. The neural net parameter setups are then proceeded and the minimal root mean square error (RMSE) is used as the performance criterion to choose a better network architecture. The RMSE is defined as follows:

$$\text{RMSE} = \frac{1}{n}\sqrt{\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{Y_i}\right)^2} \ ,$$

where n, Y_i, and $\hat{Y}_i$, are the number of samples, sample i value, and sample i prediction value, respectively. Through several pilot runs, the stopping criterion of 10,000 epochs are determined and the learning rate and the momentum coefficient are set at 0.2 and 0.9, respectively. Table 3 lists several options of the network architecture; the structure 5-5-1 (input, hidden, and output layers, respectively) is selected to obtain a better performance. The performance vs. training epochs for the network 5-5-1 is shown in Figure 2.

Table 3. Options of neural networks

Architecture	RMSE	
	Training	Testing
5-3-1	0.000002	0.215042
5-4-1	0.000002	0.215299
5-5-1	0.000001	0.193159
5-6-1	0.000003	0.201809
5-7-1	0.000002	0.203622

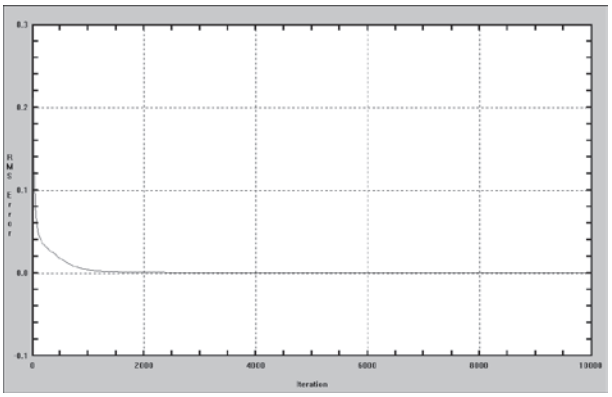

Figure 2. RMSE vs. number of training epochs

The next step of the proposed procedure estimates the missing data values through the trained network 5-5-1 and the outputs are shown as Table 4. The last step of the proposed procedure then performs Taguchi's approach to determine the optimal control factor level. The objective of the experiment is to maximize the signal/noise (S/N) ratio, η, defined as:

$$\eta = -10 \log_{10}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^2\right).$$

The resulting S/N values and factor effects are summarized in Tables 5 and 6, respectively. Note that factor B does not have a significant effect on S/N and is identified from source literature. Table 6 shows that the optimal combination is $A_2C_2D_2E_1$ which is the same as the design from complete data situation.

Table 4. Missing data estimation for example problem

No.	Control factors					Estimates
	A	B	C	D	E	
2	1	1	2	2	2	70.00115
3	1	2	1	2	2	91.66858
5	2	1	1	1	2	56.22520

Table 5. S/N values for example problem

No.	Control factors					Responses		S/N
	A	B	C	D	E			
1	1	1	1	1	1	66.0000	66.0000	− 36.3909
2	1	1	2	2	2	70.00115	63.0000	− 36.8523
3	1	2	1	2	2	91.66858	88.0000	− 38.6807
4	1	2	2	1	1	63.0000	65.0000	− 36.1247
5	2	1	1	1	2	73.0000	56.22520	− 36.6125
6	2	1	2	2	1	37.0000	42.0000	− 31.9493
7	2	2	1	2	1	38.0000	39.0000	− 31.7099
8	2	2	2	1	2	57.0000	48.0000	− 34.4350

Table 6. Factor effects for example problem

	A	B	C	D	E
Level 1	− 37.0121	− 35.4512	− 35.8485	− 35.8908	− 34.0437
Level 2	− 33.6767	− 35.2376	− 34.8403	− 34.7981	− 36.6451

In order to illustrate the solution quality of the proposed procedure, two other methods, LR and MLE, are also applied to case 1.

LR method

LR approach treats control factors and response values as independent variables and observations, respectively. The resulting regression line is:

$$y = 80.11765 - 21.80392A - 8.13725C + 16.79412E$$

Note that factors B and D are found to be nonsignificant from ANOVA results for this case. Missing data are estimated from the above equation, the S/N ratio objective are then used for completing the experimental analysis. This leads to the optimal combination of $A_2C_2E_1$.

MLE method

MLE approach is another commonly used approach for missing data analysis. We assume that the response values are normally distributed random variables. Then the likelihood equations from Anderson (1957) are used for data prediction as shown in Table 7. Inserting these estimates to Table 2, we obtain the optimal combination $A_2C_2D_2E_1$ using Taguchi's approach.

Table 7. Missing data estimation for example problem

No.	Control factors					Estimates
	A	B	C	D	E	
2	1	1	2	2	2	62.87642
3	1	2	1	2	2	86.14873
5	2	1	1	1	2	79.39948

All of the above three missing data estimation methods conclude the same result as the complete data situation for the case 1 example. In order to justify the solution quality of this proposed procedure, additional test data sets are created as follows. From Su and Miao's data set (Su and Miao 1998), cases 2 and 3 randomly remove 30% and 45% data out of the original data set, respectively. The proposed procedure is applied to both cases 2 and 3 and the results are summarized in Table 8. These results show that the proposed procedure and the other two methods obtain the same condition for the significant factors A and E. The combinations obtained by the proposed procedure are extremely close to those obtained by *MLE*. However, in case 3, the proposed procedure can be considered as a more reliable approach than the other two approaches.

Table 8. Results summary - 1

Methods	Optimal condition		
	Case 1	Case 2	Case 3
Taguchi method (Complete data)	$A_2C_2D_2E_1$ (B is not a significant factor.)		
Proposed method	$A_2C_2D_2E_1$	$A_2C_2D_2E_1$	$A_2C_2D_2E_1$
LR	$A_2C_2E_1$	$A_2C_2E_1$	$A_2D_2E_1$
MLE	$A_2C_2D_2E_1$	$A_2C_2D_2E_1$	$A_2D_2E_1$

Additional experiments adopt a larger problem of $L_{16}(2^{15})$ from Quinlan (1989) as shown in Table 9. Cases 4, 5, and 6 randomly delete 15%, 30%, and 45%, respectively, data from the complete data set (Table 9). The proposed procedure is applied to those cases and the final results are summarized in Table 10.

Table 9. Larger example data set

No.	Control factors															Responses			
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O				
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0.48	0.54	0.46	0.45
2	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	0.55	0.60	0.57	0.58
3	1	1	1	2	2	2	2	1	1	1	1	2	2	2	2	0.07	0.09	0.11	0.08
4	1	1	1	2	2	2	2	2	2	2	2	1	1	1	1	0.16	0.16	0.19	0.19
5	1	2	2	1	1	2	2	1	1	2	2	1	1	2	2	0.13	0.22	0.20	0.23
6	1	2	2	1	1	2	2	2	2	1	1	2	2	1	1	0.16	0.17	0.13	0.12
7	1	2	2	2	2	1	1	1	1	2	2	2	2	1	1	0.24	0.22	0.19	0.25
8	1	2	2	2	2	1	1	2	2	1	1	1	1	2	2	0.13	0.19	0.19	0.19
9	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	0.08	0.10	0.14	0.18
10	2	1	2	1	2	1	2	2	1	2	1	2	1	2	1	0.07	0.04	0.19	0.18
11	2	1	2	2	1	2	1	1	2	1	2	2	1	2	1	0.48	0.49	0.44	0.41
12	2	1	2	2	1	2	1	2	1	2	1	1	2	1	2	0.54	0.53	0.53	0.54
13	2	2	1	1	2	2	1	1	2	2	1	1	2	2	1	0.13	0.17	0.21	0.17
14	2	2	1	1	2	2	1	2	1	1	2	2	1	1	2	0.28	0.26	0.26	0.30
15	2	2	1	2	1	1	2	1	2	2	1	2	1	1	2	0.34	0.32	0.30	0.41
16	2	2	1	2	1	1	2	2	1	1	2	1	2	2	1	0.58	0.62	0.59	0.54

Table 10. Results summary - 2

Methods	Optimal condition		
	Case 4	Case 5	Case 6
Taguchi method (Complete data)	$A_1B_2C_2D_1E_2F_2G_2H_1I_2J_1K_1L_2M_2N_2O_2$ (A, C, D, E, F, G, H, and K are significant factors.)		
Proposed method	$A_1B_2C_2D_1E_2F_2G_2H_1I_2J_1$ $K_1L_2M_2N_2O_2$	$A_1B_2C_2D_1E_2F_2G_2H_1I_2J_1$ $K_1L_2M_2N_2O_2$	$A_1B_2C_2D_1E_2F_2G_2H_1I_2J_1$ $K_1L_2M_1N_2O_1$
LR	$A_1B_2C_2D_1E_2F_2G_2H_1$ $I_2J_1K_1L_2M_2N_2O_2$	$I_2J_1K_1L_2M_2N_2O_2$ $A_1B_2C_2D_1E_2F_2G_2H_1$	$A_1B_2C_2D_1E_2F_2G_2H_1$ $I_2J_1K_1L_2M_1N_2O_1$
MLE	$A_1B_2C_2D_1E_2F_2G_2H_1$ $I_2J_1K_1L_2M_2N_2O_2$	$I_2J_1K_1L_2M_2N_2O_2$ $A_1B_2C_2D_1E_2F_2G_2H_1$	$A_1B_2C_2D_1E_2F_2G_2H_1$ $I_2J_1K_1L_2M_2N_2O_1$

These extensive numerical experimental results show that the proposed methodology and the other two commonly used methodologies all find the same optimal condition for the significant factors as the complete data situation. The proposed methodology, therefore, is a viable approach to solve a fractional factorial experimental design problem effectively and efficiently.

4 CONCLUSIONS

This paper has presented a neural network approach for missing data estimation for the analysis of a fractional factorial experiment. Empirical results show promising for the proposed

methodology to solve a missing data fractional factorial experiment design problem. It is both effective and efficient in providing a data estimation for a variety of scenarios. The proposed methodology could be an alternative towards missing data estimations to commonly used statistical approaches since it features both the easy implementation and learning capability of a neural net.

The weakness of the proposed methodology may be from the lack of sound statistical foundations. When a user does not have sound prerequisites for using a statistical approach, the proposed methodology could be an alternative in conducting a fractional factorial experiment with missing data.

Acknowledgement

This work was supported, in part, by the Ministry of Science and Technology, Taiwan, under grant MOST-106-2221-E-006 -162 -MY3.

REFERENCES

Anderson, T. W., "Maximum likelihood estimates for the multivariate normal distribution when some observations are missing," *Journal of the American Statistical Association*, Vol. 52, pp. 200-203, 1957.

Box, G.E.P.,"Signal to noise ratios, performance criteria and transformations", *Technometrics*, Vol. 30, No. 1, pp. 1-31, 1988.

Dayhoff, J. E., *Neural Network Architectures*, Van Nostrand Reinhold, NY, 1990.

DeSarbo, W.S. and Green, P.E., "An alternating least-squares procedure for estimating missing preference data in product-concept testing", *Decision Sciences*, Vol. 17, pp. 163-185, 1986.

Enders, C. K., *Applied Missing Data Analysis*, Guilford Press, NY, 2010.

Enders, C. K., and Bandalos, D. L., "The relative performance of full information maximum likelihood estimation for missing data in structural equation models", *Structural Equation Modeling*, Vol. 8, No. 3, pp. 430-457, 2001.

Fausett, L., *Fundamentals of Neural Networks*, Prentice Hall, Englewood Cliffs, N.J., 1994.

Gleason, T.C., and Staelin, R., "A proposal for handling missing data", *Psychometrika*, Vol. 40, pp. 229-252, 1975.

Graham, J. W., "Missing data analysis: making it work in the real world", *Annual Review of Psychology*, Vol. 60, pp. 549-576, 2009.]

Kachar, R.N.,"Off-line quality control, parameter design and the Taguchi method", *Journal of Quality Technology*, Vol. 17, No. 4, pp. 176-209, 1985.

Little, R. J., and Rubin, D. B., *Statistical Analysis with Missing Data*, John Wiley & Sons, NY, 2014.

Phadke, M.S., *Quality Engineering Using Robust Design*, Prentice Hall, Englewood Cliffs, N.J., 1988.

Qnet97t *user's manual*, 1997, Vesta Services, Inc., Winnetka, IL, 1997.

Quinlan, J., "Product Improvement by Application of Taguchi Methods" , *Taguchi Methods: Applications in World Industry*, Bendell, A.D., J. Pridmore, W.A. (editors), IFS, pp. 257-266, 1989.

Ross, P.J., *Taguchi Techniques for Quality Engineering*, McGraw Hill, New York, N.Y., 1996.

Stern, H.S., "Neural networks in applied statistics", *Technometrics*, Vol. 38, pp. 205-220, 1996.

Su, C.-T. and Miao, C.-L., "Neural network procedures for experimental analysis with censored data", *International Journal of Quality Science*, Vol. 3, No. 3, pp. 239-253, 1998.

Taguchi, G., Elsayed, E.A., and Hsiang, T.C., *Quality Engineering in Production Systems*, McGraw Hill, New York, N.Y., 1989.

Wang, C.Y., Wang, S., Gutierrez, R.G., and Carroll, R.J., "Missing or truncated data - local linear regression for generalized linear models with missing data", *The Annals of Statistics*, Vol. 26, No. 3, pp. 1028-1050, 1998.