CASE 77

Mahalanobis Distance Application for Health Examination and Treatment of Missing Data

Abstract: When a Mahalanobis space is formed, there occasionally exist missing data in a group of healthy people. If these are not treated properly, the base point and unit distance cannot be determined accurately. In this study, a method to supplement the missing data was studied.

1. Introduction

In studies utilizing the Mahalanobis–Taguchi system (MTS), missing data are sometimes generated, caused by an inability to calculate a Mahalanobis distance. For example, when the inverse matrix of a correlation matrix cannot be calculated due to its multiple collinearity, and if such data are left untreated, most computer software automatically treats them as zero. As a result, calculation of an inverse matrix becomes possible, but such results are meaningless.

As a solution to this problem, the following two methods are proposed:

1. If it is possible to collect enough information with no missing data for normal people, we can create a Mahalanobis space with them.
2. Using only data for items other than those with missing data, we form a Mahalanobis space.

Method 2 is considered to be intricate and impractical because we need to recreate a Mahalanobis space for each combination of items with no missing data. In addition, although theoretically, method 2 does not require any countermeasures for missing data, accuracy in judgment could be lowered because of the decreased number of examination items.

Since we could secure no missing data for 354 normal persons in our research, we created a Mahalanobis space by adopting method 1. The main reason that the data with no missing measurements were used is that(1) we evaluated discriminability by using the data with no missing measurements, and (2) after generating missing measurements intentionally and randomly, we took effective measures and assessed the new discriminability.

Therefore, the objective of our research was to obtain a guideline for taking effective measures for missing data such that a Mahalanobis distance can be calculated accurately for a certain examinee if his or her medical checkup data were missing.

2. Collection and Sorting of Data

As our analysis data we used the data with no missing measurements in blood test items, which had been measured at the periodic medical checkups for 1377 over-40-year-old persons working at two offices of company A for three years, from February 1992 to March 1995.

Since company A's offices are distributed all over the nation, medical checkups, data processing, and final judgments are implemented at multiple medical examination facilities. If we analyzed all of the medical checkup data measured at the multiple facilities, the variability among the facilities could affect the result as a disturbance factor. Therefore, in our research we only used the data from the examination site run by company B.

Company B's comprehensive judgment on a medical checkup has the eight categories described in Table 1.

In our research, a "normal" person was defined as a person judged in the comprehensive judgment to be A_1 or A_2. Eventually, 354 people were selected as normal persons. On the other hand, for an "abnormal" person, diagnosed as C_1 or C_2, 221 people were chosen. People judged as G_1 or G_2 would have needed to retake the test or have a thorough reexamination and are not included in our study.

3. Medical Checkup Method

Using the data for 354 people matching the definition of "normal," we formed a Mahalanobis space. What is important here is which items (characteristics) to select in creating a Mahalanobis space. In general, since the results in a clinical test are largely affected by gender or age, by adding these two to the 25 items for a blood test shown in Table 2 (i.e., using a total of 27 items), we created a Mahalanobis space. In the gender category, we set data for a male to 0 and for a female to 1 and then treated the data similarly to other data.

If a Mahalanobis distance is below a "certain value," we judged it as normal and categorized it as "no guidance and precise examination are needed." If above a certain value, we categorized it as "guidance or no precise examination is needed" or "therapy is needed." To this end, we set up a threshold.

A threshold should be determined by doctors from the viewpoint of defining how far a Mahalanobis distance is positioned from a group of normal people. However, since at a medical checkup we cannot compare the health condition for each examinee with the corresponding Mahalanobis distance, we attempted to determine the threshold by using type I and type II error.

In the case of determining a threshold by the type of error, medical experts should calculate from an economic standpoint such that losses due to type I and type II error are balanced. However, how to determine the threshold using both types of error is a difficult matter. Since the objective of our research was to assess discriminability through a method of taking measures for missing data, by assuming for the sake of simplicity that the losses caused by both errors were equal, we determined the threshold.

4. Simulation of Missing Measurements and Countermeasures

To perform a simulation of measures for missing data in the medical checkup, we set up a model to generate missing data randomly for each data item, consisting of 1377 checkups for a total of 25 blood test items (Table 3). Now we defined 1, 5, 10, 20, and 30% as five levels of missing data. Following are procedures for missing-measurement simulation.

Table 1
Definition of normal and abnormal persons in comprehensive judgment

Definition	Comprehensive Judgment	Frequency	Proportion (%)
Normal persons	A_1: normal	59	4.3
	A_2: healthy with comments	295	21.4
	B_1: observation needed	345	25.1
	B_2: under observation	6	0.4
Abnormal persons	C_1: therapy needed	56	4.1
	C_2: under therapy	165	12.0
	G_1: reexamination needed	15	1.1
	G_2: precise reexamination needed	436	31.7
	Total	1377	100.1

Table 2
Blood test items used for our study

No.	Test Item	Abbreviation
1	White blood cell count	WBC
2	Red blood cell count[a]	RBC
3	Hemoglobin count[a]	Hb
4	Hematocrit count	Hct
5	Aspartate amino transferase[a]	AST
6	Alanine amino transferase[a]	ALT
7	Alkaline phosphatase	ALP
8	Glutamyltranspeptidase[a]	γ-GTP
9	Lactatdehydrogenase	LDT
10	Total bilirubin	TB
11	Thymol turbidity test	TTT
12	Zinc turbidity test	ZTT
13	Total protein	TP
14	Albumin	Alb
15	α_2-Globulins	α_2-Gl
16	β-Globulins	β-Gl
17	γ-Globulins	γ-Gl
18	Amylase	AMY
19	Total cholesterol[a]	TC
20	Triglyceride[a]	TG
21	High-density lipoprotein cholesterol	HDL-C
22	Fasting blood sugar levels	FBS
23	Blood urea nitrogen	BUN
24	Creatinine	Cr
25	Uric acid	UA

[a] Requisite. Indicates examination items that the Industrial Safety and Health Regulations (Article 44) required to be checked during a periodical medical checkup.

Table 3
Simulation model of missing data[a]

	Examination Item				
Checkup	1	2	3	···	25
1			•		•
2	•			•	
3		•			
⋮	•				•
			•		
	•			•	
1377		•			•

[a] A dot indicates missing data.

Procedure 1

Regarding the proportion of missing data as a parameter, we calculated the total number of data missing. However, since the items age and gender are not often missed, we excluded both.

Total number of missing data
= (total number of checkups)
× (number of blood test items)
× (proportion of missing data)

Procedure 2

We generated two types of random numbers:

R_1: Determine a checkup number with missing data j from all 1377 checkups.

R_2: Determine an item number with missing data i from all 25 blood test items.

If we had already generated missing data for a checkup number, j, and item number, i, we repeated procedure 2.

Procedure 3

Missing data for a checkup number, j, and item number, i, is complemented with a *complementary value,* described later.

Procedure 4

Counting the number of missing data generated, we repeated procedures 2 to 4 until the number counted reached the required total number of missing data calculated in procedure 1.

In the case of taking countermeasures for missing data, in procedure 3 we complemented a missing measurement for a checkup number, j, and item number, i, with the following three types of averages as a complementary value:

1. *Average from all examinees.* When there are missing measurements, they are quite often complemented with an average of all data because of its handiness. Our research also studied this method.
2. *Complement with item-by-item average from comprehensive judgment.* When a comprehensive judgment has been made by company B, it was expected that by complementing the data with an average from an examination item stratified by a comprehensive judgment, we could improve the discriminability more than by using an item-by-item average for all examinees.
3. *Complement with item-by-item average from normal persons.* Because a group of normal persons was regarded as homogeneous, an item-by-item average from normal persons was expected to be of significance, stable, and reliable.

This strategy is based on our supposition that examination items in which even a group of abnormal persons has missing data are likely to have almost the same values as those for a group of normal persons. Table 4 shows item-by-item averages from normal persons.

5. Results of Missing Data Calculation

Classification of Missing Data Calculation

Among the three types of countermeasures for missing data, we detailed the discriminative result in the case of complement with an item-by-item average from normal persons. Next, we compared the discriminative result in the case of leaving missing data and filling them automatically with a value of zero. For the evaluation of discriminability, an average of Mahalanobis distances for each comprehensive judgment, contribution, ρ, and SN ratio, η, were used.

Table 4
Item-by-item average from normal persons (units omitted)

Item	Examination Item	Average
1	WBC	5.85
2	RBC	476.7
3	Hb	14.93
4	Hct	44.74
5	AST	19.6
6	ALT	14.9
7	ALP	104.2
8	γ-GTP	27.3
9	LDT	343.3
10	TB	0.84
11	TTT	1.11
12	ZTT	7.27
13	TP	7.37
14	Alb	68.61
15	α_2-Gl	8.03
16	β-Gl	7.84
17	γ-Gl	12.53
18	AMY	156.9
19	TC	201.9
20	TG	94.9
21	HDL-C	58.5
22	FBS	92.6
23	BUN	16.09
24	Cr	0.87
25	UA	5.56
Number of items		354

Result of Discriminability in Case of Complementing with Item-by-Item Average from Normal Persons

Proportion of Missing Data and Fluctuation of Average of Mahalanobis Distances for Each Comprehensive Judgment Table 5 shows the fluctuation of the average of Mahalanobis distance for each comprehensive judgment when we change the proportion of the missing data in the case where they are complemented by item-by-item averages from normal persons.

When the proportion of missing data is 10% the averages of Mahalanobis distances for the two groups of normal persons, A_1 and A_2, were 1.27 and 1.60, respectively, both of which are regarded as relatively small. On the other hand, those for the two groups of abnormal persons, C_1 and C_2, were 16.11 and 6.01, respectively, which are viewed as large. Therefore, we can discriminate between normal and abnormal persons by the averages of Mahalanobis distances.

Even when the proportion of missing data is 20 or 30%, the averages of Mahalanobis distances for A_1 and A_2 are obviously smaller than those for C_1 and C_2. As a result, normal and abnormal persons can be discriminated.

Proportion of Missing Data and Change in Discriminability A 2 × 2 table (Table 6) summarizes the results of discriminability based on thresholds selected on the basis that a type I error occurs as often as a type II error does, in the case of complementing missing data with item-by-item averages from normal persons.

1. *Proportion of missing data is 1%*. When the threshold is 1.28, the occurrences of type I and type II error are 16.95% and 16.78%, respectively. The contribution, ρ, results in 43%. Since the SN ratio, η, is reduced by 0.220 dB from −1.077 dB when the proportion of missing data is zero, to −1.297 dB, the resulting discriminability decreases slightly, to 95.06% of that in the case of no missing data.
2. *Proportion of missing data is 5%*. If the threshold is set to 1.35, the occurrences of type I and type II error are 20.92 and 19.00%, respectively, whereas the contribution is 35%. The SN ratio declines by 1.605 dB, from −1.077 dB when the proportion of missing data is zero to −2.682 dB. The eventual discriminability is then lowered to 69.01% of that when there is no missing data.
3. *Proportion of missing data is 10%*. If the threshold is set to 1.55, type I and type II error take place at the possibilities of 20.92 and 19.00%,

Table 5
Average of Mahalanobis distances in case of complement with item-by-item averages from normal persons

Percent of Missing Measurements	Comprehensive Judgment							
	A_1	A_2	B_1	B_2	C_1	C_2	G_1	G_2
0	0.92	1.02	2.41	1.35	16.16	5.08	9.28	4.62
1	0.92	1.05	2.51	1.35	16.29	5.12	9.32	4.67
5	1.03	1.22	2.76	1.85	16.26	5.35	9.33	4.89
10	1.27	1.60	2.86	2.26	16.11	6.01	9.29	5.72
20	1.85	1.98	3.51	1.62	16.65	5.36	8.70	5.78
30	1.97	2.38	4.11	1.72	17.02	5.72	8.27	5.62

Table 6
Result of discriminability in case of complement with item-by-item averages for normal persons

Percent of Missing Data	Comprehensive Judgment	Mahalanobis Distance		Total	Error	ρ	η (dB)
		$D^2 \leqslant 1.26$	$D^2 > 1.26$				
0	$A_1 - A_2$	297	57	354	16.10	0.44	−1.077
	$C_1 - C_2$	37	184	221	16.74		
		$D^2 \leqslant 1.28$	$D^2 > 1.28$				
1	$A_1 - A_2$	294	60	354	16.95	0.43	−1.297
	$C_1 - C_2$	37	184	221	16.74		
		$D^2 \leqslant 1.35$	$D^2 > 1.35$				
5	$A_1 - A_2$	281	73	354	20.62	0.35	−2.682
	$C_1 - C_2$	42	179	221	19.00		
		$D^2 \leqslant 1.55$	$D^2 > 1.55$				
10	$A_1 - A_2$	283	71	354	20.06	0.36	−2.545
	$C_1 - C_2$	42	179	221	19.00		
		$D^2 \leqslant 1.80$	$D^2 > 1.80$				
20	$A_1 - A_2$	249	105	354	29.66	0.17	−6.766
	$C_1 - C_2$	61	160	221	27.60		
		$D^2 \leqslant 2.05$	$D^2 > 2.05$				
30	$A_1 - A_2$	248	106	354	29.94	0.15	−7.504
	$C_1 - C_2$	67	154	221	30.32		

respectively. Perhaps because of an uneven generation of random numbers, the resulting contribution is 35% and the SN ratio rises to −2.545 dB, up slightly from that when ρ is 5%. At the same time, since the SN ratio drops by 1.468 dB, from −1.077 dB when the proportion of missing data is zero to −2.545 dB, the discriminability diminishes to 71.32% of that when there is no missing data.

In addition, when the proportions of missing data are 20 and 30%, the SN ratios are −6.766 and −7.504 dB. These results reveal that when the proportion of missing data exceed 10%, the discriminability tends to deteriorate drastically.

Relationship between Proportion of Missing Data and Threshold

To compare the discriminability for each method of taking measures for missing data, we studied the relationship between the proportion of missing data and threshold. Figure 1 shows changes in thresholds for an increase in the number of missing data in the case where the occurrence of type I error is almost the same as that of type II error.

In the case where the proportion of missing data is 0%, if the threshold is set to 1.26, the occurrences of type I and type II error almost match. In the case of leaving missing data alone without complementing (we define this case as case 1), the threshold increases to 1.50 when the proportion of missing data is 1%, and to 6.00 when that is 5%, in a diverging manner.

On the other hand, in any case of taking measures for missing data, that is, complement with item-by-item averages from all examinees (labeled case 2), complement with item-by-item averages from comprehensive judgment (labeled case 3), and complement with item-by-item averages from normal persons (labeled case 4), the threshold increases gradually as the proportion of missing data rises.

These results demonstrates that even if we use any complementary method, the threshold becomes more stable than that in the case of leaving missing data as they are (case 1). Therefore, it is regarded as inappropriate to keep missing data intact. In addition, the threshold in the case of complement with item-by-item averages from normal persons (case 4) becomes smaller than those in the cases of complement with item-by-item averages from all

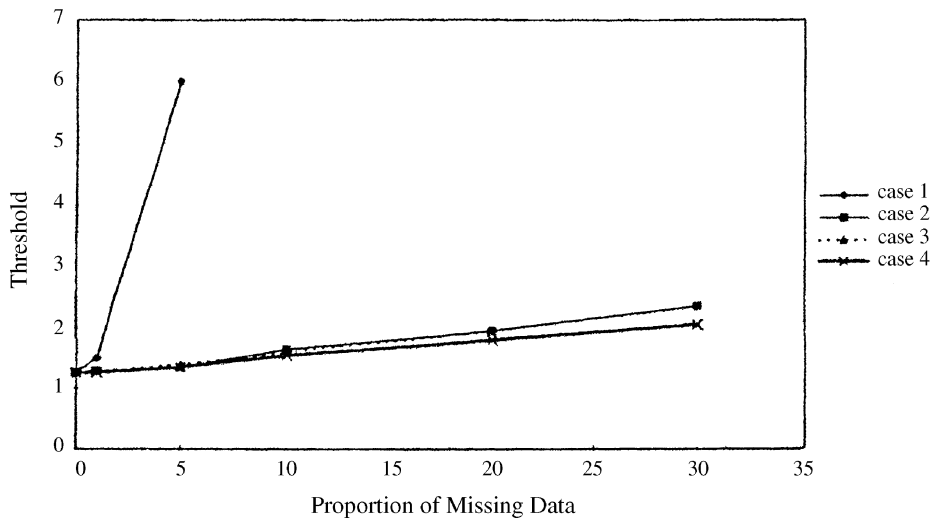

Figure 1
Proportion of missing data for each case of countermeasure and change in threshold

examinees (case 2) or with item-by-item averages from comprehensive judgment (case 3).

Relationship between Proportion of Missing Measurements and SN Ratio

Now, to compare the discriminability for each countermeasure for missing data, we studied the relationship between the proportion of missing data and the SN ratio. Figure 2 illustrates how the SN ratio varies for each measure for missing measurements as their proportion increases.

When the proportion of missing measurements is 0%, the resulting SN ratio is −1.077 dB. When missing data are kept intact (case 1), when the proportions are 1 and 5%, the corresponding SN ratios plummet drastically, to −3.548 and −10.523 dB. However, for the complement with item-by-item averages from all examinees (case 2), that of complement with item-by-item averages from each comprehensive judgment (case 3), and that of complement with item-by-item averages from normal persons (case 4), all of the SN ratios gradually decrease in a manner similar to the way the number of missing data rises.

These results indicate that no matter what complementary method is used, the SN ratio becomes more stable than when leaving missing data as they are (case 1). Comparing the discriminabilities by the SN ratio for the three complementary methods described above, complement with item-by-item averages from comprehensive judgment (case 3) is regarded as the best.

For example, when the proportion of missing measurements is 10%, the SN ratio is computed as −3.563, −2.009, or −2.545 dB for each case of complement with item-by-item averages from all examinees (case 2), complement with item-by-item averages from comprehensive judgment (case 3), and complement with item-by-item averages from normal persons (case 4). Since a larger SN ratio signifies superior discriminability, simply judging from all of the SN ratios, we can view the case of complement with item-by-item averages from comprehensive judgment (case 3) as the best.

Yet the case of complement with item-by-item averages for each comprehensive judgment (case 3) involves a problem when assuming that a comprehensive judgment can be made despite the existence of missing data. Since cases other than A_1 and A_2 in terms of a comprehensive judgment were not regarded as homogeneous, an item-by-item average from comprehensive judgment does not necessarily have a meaning. Therefore, we concluded that complement with item-by-item averages from comprehensive judgment (case 3) is not a proper method.

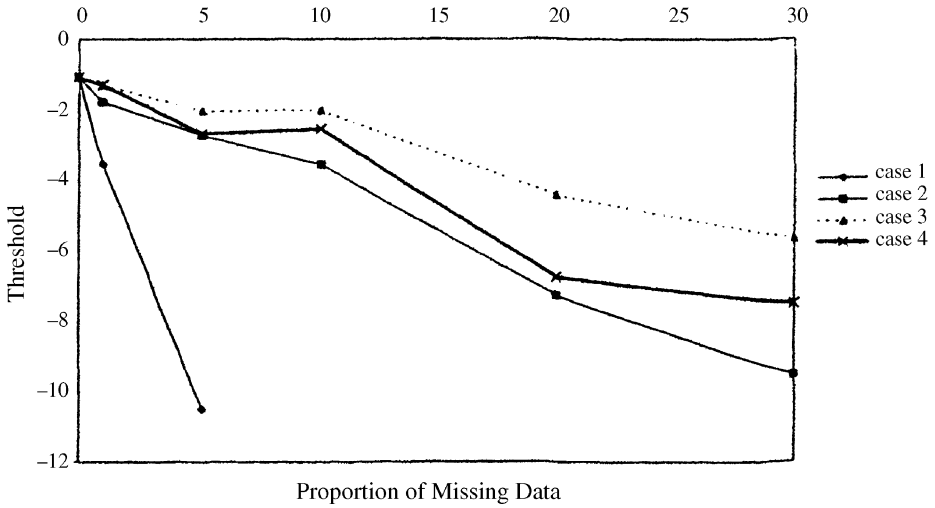

Figure 2
Proportion of missing data for each case of countermeasure and change in SN ratio

On the other hand, we compared cases of complement with item-by-item averages from all examinees (case 2) and complement with item-by-item averages from normal persons (case 4). The SN ratio for complement with item-by-item averages from normal persons (case 4) shows a better value (i.e., better discriminability) than complement with item-by-item averages for normal persons (case 4) does.

Since an item-by-item average for all examinees denotes a mean value for all normal, abnormal, and even other-type persons, its stability depends greatly on the constitution of data collected. In contrast, because a group of normal persons is homogeneous, an item-by-item average from normal persons can be considered stable.

Considering all of the above, as the most appropriate method of complementing missing data for medical checkups, we selected the complement with item-by-item averages from normal persons (case 4).

6. Countermeasures for Missing Data

Missing Data Occurrence and Simulation Model

In our research, we performed a simulation based on a model that generates missing measurements equally and randomly for any examination item. However, the actual occurrence of missing data for each examination item for medical checkups is not uniform but is rather low for the particular requisite items prescribed by Industrial Safety and Health Regulations (Table 2). This implies that important items tend to have a smaller number of missing data.

In addition, when we complemented missing measurements with item-by-item averages from normal persons (case 4), because an item-by-item average from normal persons was used even for abnormal persons' missing data, the threshold resulted in a smaller value than for complementing with other types of averages (cases 2 and 3). The reason is likely to relate to our setup of a simulation model that randomly generates missing data for the 25 blood test items.

Taking all of the above into account, we can observe that by setting up a simulation model that creates missing data in accordance with an item-by-item proportion of actual missing data, we can properly complement missing data with an item-by-item average from normal persons.

Setup of Threshold

For the sake of convenience, our research determined thresholds on the assumption that type I and type II error have the same loss. Although a threshold should essentially be computed from the loss function, it is not easy to determine a threshold that balances future losses of type I and type II error. Thus, focusing on our objective of taking effective measures for missing data for medical checkups, on the assumption that both errors' losses are equal, we determined thresholds. Since the issue of threshold selection is regarded as quite crucial, we continue to work on this issue.

7. Conclusions

In this study of a medical checkup method based on a Mahalanobis distance, we studied a calculation method for estimating a Mahalanobis distance accurately for each examinee in the case where there were missing data for medical checkups, and finally, obtained the following results.

1. When we kept missing data intact, and eventually most computer software complemented them automatically with a value of zero, even if the proportion of missing data was only 1 or 5%, the average of the Mahalanobis distances for each comprehensive judgment diverges, preventing us from discriminating between normal and abnormal persons. Therefore, it was regarded as inappropriate to leave missing data untreated.
2. The method of complementing missing data with item-by-item averages from all examinees has better discriminability than that of keeping missing data intact. However, different constitution of data collected often causes instability of averages.
3. Discriminability by the method of complementing missing data with item-by-item averages from comprehensive judgment is viewed as the best of all from the perspective of the

SN ratio. However, a group of abnormal persons is so heterogeneous that its average hardly has significance. Thus, if the method of complementing missing data with item-by-item averages from comprehensive judgment is used for cases other than A_1 and A_2, its use is not considered appropriate.

4. In terms of the SN ratio, discriminability by the method of complementing missing data with item-by-item averages from normal persons turns out to be better than the case of complement with item-by-item averages from all examinees. The reason is that because a group of normal persons is homogeneous enough that its average is regarded as meaningful, and because missing data are quite unlikely to occur in essential examination items that are supposed to distinguish between normality and abnormality, viewing missing data as normal values is not considered to distort the actuality.
5. Taking into consideration all of these matters, we judged that the method of complementing missing measurements with item-by-item averages from normal persons was the best of the measures for missing measurements.

The greatest significance in our study is that we have successfully elucidated the possibility of retaining medical checkup discriminability resting on the Mahalanobis distance with a minimal loss of information by complementing with item-by-item averages from normal persons when missing data for medical checkups. The result gained from our research is believed to contribute to the rationalization of a medical checkup: that is, reduction in medical expenses through the elimination of excessive detailed examinations, mitigation of the time and economic cost to examinees, alleviation of examinee anxiety, or assistance with information or instructions from doctors, nurses, or health workers.

Pattern recognition using a Mahalanobis distance can be applied not only to a medical checkup but also to a comprehensive judgment in many fields. Our research could contribute to the problem of missing measurements in any application.

References

Yoshiko Hasegawa, 1997. Mahalanobis distance application for health examination and treatment of missing data: 1. The case with no missing data. *Quality Engineering*, Vol. 5, No. 5, pp. 46–54.

Yoshiko Hasegawa, 1997. Mahalanobis distance application for health examination and treatment of missing data: 2. The case with no treatment and substituted zeros for missing data. *Quality Engineering*, Vol. 5, No. 6, pp. 45–52.

This case study is contributed by Yoshiko Hasegawa.