Discrimination Model for Synthetic Variables Generated from Explanatory Variables Considering Sample Attributes

Haruka Yamashita †
School of Creative Science and Engineering, Waseda University.
3-4-1, Okubo, Shinjuku-ku, Tokyo 169-8555, JAPAN
Tel:(+81) 3-5286-3290, Email: h.yamashita@aoni.waseda.jp

Masayuki Goto†
School of Creative Science and Engineering, Waseda University.
3-4-1, Okubo, Shinjuku-ku, Tokyo 169-8555, JAPAN
Tel:(+81) 3-5286-3290, Email: masagoto@waseda.jp

Abstract. Discrimination analysis is a well-known technique applicable to various kinds of problems in order to know the structural difference of explanatory variable coefficients between categories. In real data analysis, there are cases that explanatory variables of data form synthetic variables; however, the effect of the synthetic variables for the decision of the category is unknown. Moreover, there are also several cases that the effects of synthetic variables are different depending on the sample attributes. For example, we want to know the effect of factors whether a company in service industry feels successful or not and the factors (Financial, Customer, Internal Business, and Innovation and Learning indicators) that cannot be observed directly. However, synthesizing the questionnaire survey answers enables to score each factor, and the effects of the factors are different depending on the attribute of each company. In this case, the discrimination model for the factors considering the difference of the attribute of the each company is reasonable. In this study, we focus on such situation and we propose a discrimination model for synthetic variables generated from explanatory variables considering the difference of sample attributes. We verify the effectiveness of our proposed method by analyzing real-world data.

Keywords: Discrimination analysis, Sample attributes, Principal component analysis, Multivariate analysis

1. INTRODUCTION

Discrimination analysis, originally, proposed by Fisher (1936) is a well-known technique applicable to various kinds of problems in order to know the structural difference of explanatory variable coefficients between categories. Many models related to the discrimination analysis model are proposed (e.g., Cai and Liu, 2012; Bouveyron, et al., 2015) and there are rich examples of the data analysis by applying discrimination model to real data (e.g., Laddi, et al., 2013; Gerpott, et al., 2015).

In the real-world data analysis, there are cases that data is known to have hierarchical structure and synthetic variables are generated by given explanatory variables; however, the effect of the synthetic variables for the discrimination is unknown. Moreover, there are also several cases that the effects of synthetic variables are different depending on the sample attributes.

For example, there is the questionnaire data (Mizuno and Suzuki, 2010) of the managers' opinion whether the performance of the companies successful or not that is designed for knowing the effects of the four factors (Financial, Customer, Internal Business, and Innovation and Learning indicators) that are not observed directly; however, we can assume that the factors can be scored by synthesizing some answers of the questionnaire. Moreover, it is expected that the effect of the factors for the discrimination whether the company feels successful or not is different depending on the employment system of the company. Note that the data of companies was not analyzed by linear discriminant analysis, but the network analysis and see the relationship between each query (Mizuno and

Suzuki, 2010).

A natural analysis for this typed data (Zhao, et al., 1998) is that:

(1) Stratify the samples into groups by the sample attributes.
(2) Generate synthesized variables of each sample attribute by the principal component analysis (Hotteling, 1933) for corresponding explanatory variables of each sample attribute.
(3) Each first principal component score is regarded as the synthesized variables of each sample attribute and the linear discrimination models using the synthesized variables of each sample attribute are estimated for each stratified group.

However, this approach has two problems:

(a) The synthesized variables generated for each sample attribute are independent of the discrimination of the data.
(b) Since the synthesized variables are generated for each attribute, it is difficult to compare the effects for the discrimination of each synthesized variables between the different attributes.

Therefore, a method that overcomes the two problems mentioned above is required. In this study, we focus on such situation and we propose a discrimination model for synthetic variables made from explanatory variables considering sample attributes.

Moreover, we verify the effectiveness of our proposed method by analyzing real-world data. In the example, we apply the case problem to discriminate whether a baseball player was selected to the all-star game in 2015 or not in Japanese pro-baseball league. Several kinds of annual hit records of each player in 2015 are used as the explanatory variables, and the explanatory variables can be synthesized for the factors "power hitter", "stable hitter" and "speedy hitter" considering the binary attribution "the player is Japanese or foreigner".

Table 1: Data description for discriminant analysis

	1	⋯	i	⋯	I	Class
1	x_{11}	⋯	x_{1i}	⋯	x_{1I}	C_1
2	x_{21}	⋯		⋯	x_{2I}	C_2
⋮	⋮	⋮		⋮	⋮	⋮
⋮	⋮		⋮		⋮	⋮
⋮	⋮	⋮		⋮	⋮	⋮
n	x_{n1}	⋯	x_{ni}	⋯	x_{nI}	C_2
⋮	⋮	⋮		⋮	⋮	⋮
⋮	⋮		⋮		⋮	⋮
⋮	⋮	⋮		⋮	⋮	⋮
N	x_{N1}	⋯	x_{Ni}	⋯	x_{NI}	C_1

The reminder of this paper is as follows. In section 2, we describe the preliminaries of this study; the discriminant analysis and the statement of the problem of this study. In section 3, we propose a discrimination model for synthetic variables generated from explanatory variables considering score parameters of sample attributes. In section 4, we verify the effectiveness of our proposed method by analyzing real-world data using our method. Then the conclusion is described in section 5.

2. PRELIMINARIES

2.1 Linear discrimination analysis

Linear discrimination analysis, originally, proposed by Fisher (1936) is a classical statistical method. This method is applicable to various kinds of problems in order to know the structural difference of explanatory variable coefficients between categories.

Here, discrimination analysis is stated following the definition (Fisher, 1936). Let i be an explanatory variable (i=1,..., I), and let C_1 and C_2 be the classes of the each data indexed by n (n=1,...,N). The setting is figured in Table 1.

Linear discrimination analysis estimates the linear function F_n that separates classes the best.

$$F_n = \sum_{i=1}^{I} a_i x_{ni} + a_0 \quad , \qquad (1)$$

where we assume that the data vector $\boldsymbol{x}_n = (x_{n1}, \dots, x_{nI})$ $(n = 1, \dots, N)$ is drawn from a normal distribution, $a_i (i = 1, \dots, I)$ denotes the coefficients of each explanatory variables and a_0 denotes the interception of the function. The optimal parameters are estimated such that the ratio of the group-between variance to the group-within variable is maximized (Fisher, 1936). The class of $\boldsymbol{x}_n$ is determined by the value of F_n as follows:

$$\begin{cases} F_n > 0 \Rightarrow C_1 \\ F_n \leq 0 \Rightarrow C_2 \end{cases}$$

There are numerous examples of the data analysis by applying the discrimination analysis to real data (e.g., Laddi, et al., 2013; Gerpott, et al., 2015).

2.2 Assumed situation of this study

In the real-world data analysis, there are cases that a structure that explanatory variables form synthetic variables is already known; however, the effect of the synthetic variables for the decision of the category is unknown.

Table 2: Assumed data structure of this study

	i	1					⋯	i					⋯	I					Class
k	s(k)\j(i)	1(1)	⋯	j(1)	⋯	J(1)	⋯	1(i)	⋯	j(i)	⋯	J(i)	⋯	1(I)	⋯	j(I)	⋯	J(I)	
1	1(1)	$x_{1(1)1(1)}$	⋯	$x_{1(1)j(1)}$	⋯	$x_{1(1)J(1)}$	⋯	$x_{1(1)1(i)}$	⋯	$x_{1(1)j(i)}$	⋯	$x_{1(1)j(i)}$	⋯	$x_{1(1)1(I)}$	⋯	$x_{1(1)j(I)}$	⋯	$x_{1(1)j(I)}$	C_1
	⋮	⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮	⋮
	s(1)	$x_{s(1)1(1)}$	⋯	$x_{s(1)j(1)}$	⋯	$x_{s(1)J(1)}$	⋯	$x_{s(1)1(i)}$	⋯	$x_{s(1)j(i)}$	⋯	$x_{s(1)j(i)}$	⋯	$x_{s(1)1(I)}$	⋯	$x_{s(1)j(I)}$	⋯	$x_{s(1)j(I)}$	C_2
	⋮	⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮	⋮
	S(1)	$x_{S(1)1(1)}$	⋯	$x_{S(1)j(1)}$	⋯	$x_{S(1)J(1)}$	⋯	$x_{S(1)1(i)}$	⋯	$x_{S(1)j(i)}$	⋯	$x_{S(1)j(i)}$	⋯	$x_{S(1)1(I)}$	⋯	$x_{S(1)j(I)}$	⋯	$x_{S(1)j(I)}$	C_1
2	1(2)	$x_{1(2)1(1)}$	⋯	$x_{1(2)j(1)}$	⋯	$x_{1(2)J(1)}$	⋯	$x_{1(2)i1}$	⋯	$x_{1(2)j(i)}$	⋯	$x_{1(2)j(i)}$	⋯	$x_{1(2)1(I)}$	⋯	$x_{1(2)j(I)}$	⋯	$x_{1(2)j(I)}$	C_1
	⋮	⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮	⋮
	s(2)	$x_{s(2)1(1)}$	⋯	$x_{s(2)j(1)}$	⋯	$x_{s(2)J(1)}$	⋯	$x_{s(2)1(i)}$	⋯	$x_{s(2)j(i)}$	⋯	$x_{s(2)j(i)}$	⋯	$x_{s(2)1(I)}$	⋯	$x_{s(2)j(I)}$	⋯	$x_{s(2)j(I)}$	C_2
	⋮	⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮		⋮	⋮
	S(2)	$x_{S(2)1(1)}$	⋯	$x_{S(2)j(1)}$	⋯	$x_{S(2)J(1)}$	⋯	$x_{S(2)1(i)}$	⋯	$x_{S(2)j(i)}$	⋯	$x_{S(2)j(i)}$	⋯	$x_{S(2)1I}$	⋯	$x_{S(2)j(I)}$	⋯	$x_{S(2)j(I)}$	C_1

Moreover, there are also several cases that the effects of synthetic variables are different depending on the sample attributes. In this study, we propose a linear discriminant model for the synthetic variables made from explanatory variables considering the difference of sample attributes. In this subsection, we state the assumed situation of this study with introducing an example of the discrimination of the batting data of baseball players in Japanese professional baseball league.

Firstly, let i $(i = 1, \dots, I)$ be synthesized variables (e.g., the factors of "power hitter" and "stable hitter"), let $j(i)$ $(j = 1(1), \dots, J(I))$ be explanatory variables (e.g., the values of "the number of home run", "the number of three base hit", "hit ratio", "number of hits", and "on-base percentage"), let k $(k = 1, 2)$ be an attribute of sample (e.g., Japanese or foreign player), and let $s(k)$ $(s(k) = 1(1), \dots, S(1), 1(2), \dots, S(2))$. We assume that the value of "power hitter" is generated from the explanatory variables "the number of home run", "the number of three base hit", and the value of "stable hitter" is made from the explanatory variables "hit ratio", "number of hits", and "on-base percentage". Then we construct a discriminant model to classify whether a baseball player was selected to all-star game or not (i.e., the class is C_1 or C_2) considering the difference that the player is Japanese or not. The structure of assumed data frame is figured in Table 2.

The natural approach of the analysis (Zhao, et al., 1998) is to make the values of synthetic variables $z_{s(k)i}$ based on the principal analysis (using the principal score) of the values of corresponding explanatory variables for each sample attribute $x_{s(k)j(i)}$, then construct the discriminant model of classes . C_1 or C_2 .

However, this approach makes synthesized variables of each attribute without considering the discrimination. There are two problems of this approach:

(a) By the approach for making the values of synthesized variables $z_{s(k)i}$ for each sample attribute k, the coefficient parameters of synthesized variables are calculated independently of the problem of discrimination.
(b) Since the values of synthesized variables $z_{s(k)i}$ are made for each sample attribute k, the comparison of the effects for the discrimination of each synthesized variables between each attribute is difficult.

3. PROPOSED METHOD

3.1 Discrimination Model for Synthetic Variables Considering Sample Attributes

In this study, we propose a discriminant model for synthetic variables generated from explanatory variables considering sample attributes.

Let $b_{j(i)}$ be a coefficient of the explanatory variables for the discrimination, and a_{ki} be also a coefficient of the synthetic variables for the discrimination. Then, the discrimination function $F_{s(k)}$ is defined as follows:

$$F_{s(k)} = \sum_{i=1}^{I} a_{ki} \sum_{j(i)=1(i)}^{J(i)} b_{j(i)} x_{s(k)j(i)} + c_0, \qquad (2)$$

where we assume that the $\sum_{i=1}^{I}\sum_{j(i)=1(i)}^{J(i)} j(i)$ dimensional explanatory variables vector $\boldsymbol{x}_{s(k)}=(x_{s(k)j(i)})$ is drawn from a normal distribution, and c_0 denotes the interception of the function. In the model, $\sum_{j(i)=1(i)}^{J(i)} b_{j(i)} x_{s(k)j(i)}$ is assumed to be the generation of synthetic variables for a linear discriminant function $F_{s(k)}$. The a_{ki} in equation (2) is the coefficient of the discrimination for the synthesized variables for each sample attribute k. The class is determined by the value of $F_{s(k)}$ as follows:

$$\begin{cases} F_{s(k)} > 0 \Rightarrow C_1 \\ F_{s(k)} \leq 0 \Rightarrow C_2 \end{cases}$$

In this model, there are mainly two advantages.

(a) When we synthesize the explanatory variables, we do not use principal component analysis but optimize the parameter $b_{j(i)}$. Therefore, our synthesizing method of the explanatory variables considers the discrimination.
(b) We do not optimize parameter of each explanatory vector of each sample parameter but we optimize $b_{j(i)}$ commonly between the different sample attributes. This approach enables to compare the estimated parameter of a_{ki} for each sample attribute k.

The two advances overcome the two problems represented in section 2.2, respectively.

3.2 Parameter estimation of the discriminant function

Since the parameters a_{ki} and b_{ij} are not optimized by the analytic approach, we estimate the parameters a_{ki} and b_{ij} alternately as follows:

STEP1 (Initialization)

Set the all values of $a_{ki} = 1$.

STEP2 (Parameter estimation of $\boldsymbol{b_{ij}}$)

The a_{ki} are settled and calculate each

$$y_{s(k)j(i)} = a_{ki} x_{s(k)j(i)}.$$

Then, equation (2) is written as follows;

$$F_{s(k)} = \sum_{i=1}^{I} \sum_{j(i)=1}^{J(i)} b_{j(i)} y_{s(k)i} + c_0,$$

Here, let $\bar{\boldsymbol{y}}^{[1]} = (\bar{y}_i^{[1]}), \bar{\boldsymbol{y}}^{[2]} = (\bar{y}_i^{[2]})$ be $\sum_{i=1}^{I}\sum_{j(i)=1(i)}^{J(i)} j(i) = \alpha$ dimensional estimator of the population mean of $y_{s(k)i}$ in class C_1,C_2, respectively, and $\boldsymbol{S}$ be the estimator of $\alpha \times \alpha$ population covariance matrix of $y_{s(k)i}$. The vector $\boldsymbol{b} = (b_{ji})$ is estimated as

$$\boldsymbol{b} = \boldsymbol{S}^{-1} \ (\bar{\boldsymbol{y}}^{[1]} - \ \bar{\boldsymbol{y}}^{[2]}).$$

Then, $\boldsymbol{b}$ is standardized such that the sum of the squared value is 1.

STEP3 (Parameter estimation of $\boldsymbol{a_{ki}}$)

Settle the estimated $\boldsymbol{b}$ in STEP2, and let

$$z_{s(k)i} = \sum_{j(i)=1(i)}^{J(i)} b_{j(i)} x_{s(k)j(i)}.$$

Then, equation (2) is written as follows;

$$F_{s(k)} = \sum_{i=1}^{I} a_{ik} z_{s(k)i} + c_0,$$

Here, let $\bar{\boldsymbol{z}}^{[1]} = (\bar{z}_{ki}^{[1]}), \bar{\boldsymbol{z}}^{[2]} = (\bar{z}_{ki}^{[2]})$ be the estimator of $2I$-dimensional population mean of $z_{s(k)i}$ considering the difference of sample attribute k in classes C_1, C_2, respectively, and $\boldsymbol{S'}$ be the $2I \times 2I$ population covariance matrix. The vector $\boldsymbol{a} = (a_{kj})$ is estimated as

$$\boldsymbol{a} = \boldsymbol{S'^{-1}} \ (\bar{\boldsymbol{z}}^{[1]} - \ \bar{\boldsymbol{z}}^{[2]}).$$

STEP4 (Judge of the conversion)

Iterate STEPS 2 and 3 until the parameters do not change.

STEP5 (Parameter Estimation of $\boldsymbol{c_0}$)

Estimate the intercept of the discriminant function as follows:

$$c_0 = \boldsymbol{a'} \ (\bar{\boldsymbol{z}}^{[1]} + \ \bar{\boldsymbol{z}}^{[2]}).$$

Note that, since this algorithm introduces the alternative

approach, the estimator $\boldsymbol{b}$, $\boldsymbol{a}$, and c_0 may be the local optimal (not always global optimal). Also, another approach for estimating $\boldsymbol{b}$ is to apply the principal components analysis and let $z_{s(k)i}$ be the each first principal component scores calculated from the corresponding explanatory variables; however, this approach cannot take account of the discrimination of the classes, then our method is reasonable in terms of the data discrimination.

4. ANALYSIS EXAMPLE

4.1. Experimental conditions

In this section, we verify the effectiveness of our proposed method by analyzing real-world data. In the example, we use the data of discrimination whether the baseball player was selected to the all-star game or not in Japanese professional baseball league in 2015.

In Japan, professional baseball league holds an annual event "all-star games". The players are selected in terms of their records and popularity. We selected 58 players who achieve to Provisions at-bat in 2015*. We made the data that includes 8 explanatory variables for the three synthetic variables "Power hitter", "Stable hitter", and "Fast hitter" considering that the player is Japanese or Foreigner.

Table 2: Explanatory variables and synthesized variables

Synthesized	Power hitter		
Explanatory	# of home run	% of long hit	hit point
Synthesized	Stable hitter		
Explanatory	% of hit	# of hit	% of on-base
Synthesized	Fast hitter		
Explanatory	# of base steal	# of 3 base hits	

Table 3: The discrimination rate of the analysis

(i) Zhao et al. (1998)	(ii) Proposal
63.50%	76.90%

The detailed description of explanatory variables and synthesized variables is in Table 2. Note that each explanatory score are standardized.

We analyzed the data by two approaches (i) the method (Zhao et al.,1998), and (ii) our proposal, respectively.

4.2. Results and interpretations

The results of the analysis for the discrimination rate of two methods are shown in Table 3. This result shows that although the number of parameters is smaller, our proposed method represents the better result than the result obtained by the conventional method (Zhao, et al. 1998).

Let us describe the interpretation of the obtained results comparing two approaches. Since the conventional method estimates the score of each explanatory variable of each sample attribute, the comparison of the parameter of each synthetic variable between each sample attribute is difficult. On the other hand, in our method, the scores of each explanatory variable are common between the sample attributes; the comparison of the parameter of each synthetic variable for each sample attribute is available. Therefore, we can interpret the results easily.

Table 4: Estimated parameters obtained by the conventional method (Zhao, et al. 1998)

Japanese players							
Power hitter			Stable hitter			Fast hitter	
0.36			0.24			0.23	
Home run	Long hit	Hit point	% of hit	# of hit	% of on-base	Base steal	3 base hit
0.58	0.58	0.57	0.61	0.59	0.54	0.67	0.74
Foreign players							
Power hitter			Stable hitter			Fast hitter	
0.36			0.24			0.54	
Home run	Long hit	Hit point	% of hit	# of hit	% of on-base	Base steal	3 base hit
0.67	0.49	0.55	0.53	0.26	0.8	0.32	0.95

* Pro-baseball freak (http://baseball-freak.com/)

Japanese players							
Power hitter			Stable hitter			Fast hitter	
0.30			0.66			0.14	
Foreign players							
Power hitter			Stable hitter			Fast hitter	
0.43			0.20			0.49	
Home run	Long hit	Hit point	% of hit	# of hit	% of on-base	Base steal	3 base hit
-0.54	0.53	0.30	-0.31	0.40	-0.05	0.30	-0.07

Especially, focusing on the scores of the explanatory variables for the "stable hitter", since the scores of Japanese players are clearly different from the scores of foreigner players, the same value of the parameter of "stable hitter" of both Japanese players and foreigner players are difficult to be interpreted. Therefore, our method which represents the scores of the explanatory variables commonly between each sample attributes, shows the difference of the effect of "stable hitter" for the discrimination between Japanese and foreigner better. Moreover considering other results, they are fit to our experimental knowledge.

These results suggested that our model is reasonable in terms of both the accuracy and the interpretation.

5. CONCLUSION AND FUTURE WORKS

In this study, we focus on the data whose explanatory variables form synthetic variables and the effects of synthetic variables are different depending on the sample attributes, and we propose a discrimination model for synthetic variables made from explanatory variables taking account of the differnce of sample attributes.

We also verify the effectiveness of our proposed method by analyzing real-world data of synthetic variables generated from explanatory variables considering sample attributes using our method. The results suggested the advantages of our proposal.

However, we have demonstrated the analysis only for one data set, and for confirming the usability, we have to analyze more datasets. Also, there is no consideration for variables selection. In fact, in the real-world data, there are many cases that some explanatory variables have strong co-relation or a variable does not contribute to discriminate the categories. Therefore, the approach considering the variable selection is required.

ACKNOWLEDGMENTS

A part of this study was supported by JSPS KAKENHI Grant Numbers 26282090, 26560167, and 16K16361.

REFERENCES

Bouveyron, C., Fauvel, M., and Girard, S.. (2015), Kernel discriminant analysis and clustering with parsimonious Gaussian process models, *Statistics and Computing*, **25**, 1143―1162.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. Annals of Eugenics, **7**, 179–188.

Gerpott, T., Ahmadi, N., and D Weimar, D. (2015), Who is (not) convinced to withdraw a contract termination announcement?–A discriminant analysis of mobile communications customers in Germany, *Telecommunications Policy*, **39**, 38---52

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology,* **24**, 417–441.

Kei, Mizuno and Hideo, Suzuki. (2010), A Study on Relationships Sports Clubs between Performance of Professional Using Network Analysis, *Journal of Japan Industrial Management Association*, **61**, 263-274.

Laddi, Amit., Neelmam Rup Prakash, Shashi Sharma, and Amod Kumer (2013), Discrimination analysis of Indian tea varieties based upon color under optimum illumination, *Journal of food measurement & characterization*, **7**, 60―65.

Tony Cai and Weidong Liu (2011), A Direct Estimation Approach to Sparse Linear Discriminant Analysis, *Journal of the American Statistical Association*, **106**, 1566--1577.

Zhao Wenyi, Arvindh, Krishnaswamy, Chellappa Rama, Swets Daniel L., Weng John, (1998). "Discriminant Analysis of Principal Components for Face Recognition," F ace Recognition, Vol.163, pp.73–85, 1998.