Four Serious Problems and New Facts of the Discriminant Analysis

Shuichi Shinmura(✉)

Faculty of Economics, Seikei University, Tokyo, Japan
shinmura@econ.seikei.ac.jp

Abstract. The discriminant analysis is essential knowledge in science, technology and industry. But, there are four serious problems. These are resolved by Revised IP–OLDF and k–fold cross validation.

Keywords: Minimum number of misclassifications (MNM) · Revised IP-OLDF · SVM · LDF · Logistic regression · k-fold cross validation

1 Introduction

Fisher [1] described the linear discriminant function (**LDF**), and founded the discriminant theory. Following this, the quadratic discriminant function (**QDF**) and multi-class discrimination using Mahalanobis distance were proposed. These functions are based on the variance-covariance matrices, and are easily implemented in the statistical software packages. They can be used in many applications. However, real data rarely satisfy Fisher's assumptions. Therefore, it is well known that logistic regression is better than LDF and QDF, because it does not assume a specific theoretical distribution, such as a normal distribution. In addition to this, the discriminant rule is very simple: If $y_i*f(\mathbf{x_i}) > 0$, $\mathbf{x_i}$ is classified to class1/class2 correctly. If $y_i*f(\mathbf{x_i}) < 0$, $\mathbf{x_i}$ is misclassified. There are four serious problems hidden in this simplistic scenario [22].

(1) Problem 1

We cannot properly discriminate between cases where $\mathbf{x_i}$ lies on the discriminant hypeplane ($f(\mathbf{x_i}) = 0$). This **unresolved problem** has been ignored until now. The proposed **Revised IP-OLDF** is able to treat this problem appropriately. Indeed, except for Revised IP-OLDF, no functions can correctly count the number of misclassifications (NM). These functions should count the number of cases where $f(\mathbf{x_i}) = 0$, and display this alongside the NM in the output.

(2) Problem 2

Fisher's LDF and QDF cannot recognize linear separable data (where the Minimum NM (MNM) = 0). This fact was first found when **IP-OLDF** was applied to Swiss bank note data [3]. In this paper, the determination of pass/fail in exams is used because it is trivially linear-separable and we can obtain it easily. We show that, in many cases, the NMs of LDF and QDF are not zero. Next, 100 re-samples of these data are generated, and the mean error rates are obtained by 100-fold cross validation. The mean error rates

E. Pinson et al. (Eds.): ICORES 2014, CCIS 509, pp. 15–30, 2015.
DOI: 10.1007/978-3-319-17509-6_2

of LDF are 6.23 % higher than that of Revised IP-OLDF in the validation samples of Table 7.

(3) **Problem 3**

If the variance-covariance matrix is singular, Fisher's LDF and QDF cannot be calculated because the inverse matrices do not exist. The LDF and QDF of JMP [9] are solved by the generalized inverse matrix technique. In addition to this, **RDA** [4] is used if QDF causes serious trouble with dirty data. However, RDA and QDF do not work properly for the special case in which the values of features belonging to one class are constant. If users can choose proper options for a **modified RDF** developed for this special case, it works better than QDF and LDF in Table 5.

(4) **Problem 4**

Some statisticians misunderstand that the discriminant analysis is the inferential statistical method as same as the regression analysis, because it is derived from Fisher's assumption. But there are no standard error (SE) of the discriminant coefficients or error rate, and variable selection methods such as stepwise methods and statistics such as Cp and AIC. In this paper, we propose "k-fold cross validation for small samples" and new variable selection method, the minimum mean error rates of which is chosen as the best model. In future works (**Future1**), generalization ability and 95 % confidence intervals of all LDFs are proposed.

In this research, two Optimal LDFs (**OLDFs**) based on the MNM criterion are proposed. The above three problems are solved by **IP-OLDF** and **Revised IP-OLDF** completely. IP-OLDF [13–15] reveals the following properties.

Fact (1) Relation between LDFs and NMs. IP-OLDF is defined on the data and discriminant coefficient spaces. Cases of $\mathbf{x_i}$ correspond to linear hyper-planes ($H_i(\mathbf{b}) = y_i{*}\ ({}^t\mathbf{x_i}\mathbf{b} + 1) = 0$) in the p-dimensional discriminant coefficient space that divide the space into two half-planes: the plus half-plane ($H_i(\mathbf{b}) > 0$) and minus half-plane ($H_i(\mathbf{b}) < 0$). Therefore, the coefficient space is divided into a finite convex polyhedron by $H_i(\mathbf{b})$. Interior point $\mathbf{b}_j$ of the convex polyhedron corresponds to the discriminant function $f_j(\mathbf{x}) = {}^t\mathbf{b}_j\mathbf{x} + 1$ on the data space that discriminates some cases properly and misclassifies others. This means that each interior point $\mathbf{b}_j$ has a unique NM. The "**Optimal Convex Polyhedron (OCP)**" is defined as that with the MNM. Revised IP-OLDF [16] can find the interior point of OCP directly, and solves the unresolved problem (**Problem 1**) because there are no cases on the discriminant hyper-plane ($f(\mathbf{x_i}) = 0$). If $\mathbf{b}_j$ is on a vertex or edge of the convex polyhedron, however, the unresolved problem cannot be avoided because there are some cases on $f(\mathbf{x_i}) = 0$.

Fact (2) Monotonous decrease of MNM ($\mathbf{MNM_p \geq MNM_{(p+1)}}$). Let MNM_p be the MNM of p features (independent variables). Let $MNM_{(p+1)}$ be the MNM of the (p + 1) features formed by adding one feature to the original p features. MNM decreases monotonously ($MNM_p \geq MNM_{(p+1)}$), because OCP in the p-dimensional coefficient space is a subset of the (p + 1)-dimensional coefficient space [18]. If $MNM_p = 0$, all MNMs including p features are zero. Swiss bank note data consists of genuine and counterfeit bills with six features. IP-OLDF finds that this data is linear-separable according to two features (X4, X6). Therefore, 16 models including these two features have MNMs = 0. Nevertheless, Fisher's LDF and QDF cannot recognize that this data is linear-separable, presenting a serious problem. In this paper, we show that Revised IP-OLDF can resolve the above three problems, and is superior to Fisher's LDF, logistic regression, and Soft-margin SVM (S-SVM) [28] under 100-fold cross validation [20, 21] of the pass/fail determinations of exams [19] and their re-sampled data.

2 Discriminant Functions

2.1 Statistical Discriminant Functions

Fisher defined LDF to maximize the variance ratio (between/within classes) in Eq. (1). This can be solved by non-linear programming (NLP).

$$\mathrm{MIN} = {}^{t}\mathbf{b}(\mathbf{m}_1 - \mathbf{m}_2)^{t}(\mathbf{m}_1 - \mathbf{m}_2)\mathbf{b}/{}^{t}\mathbf{b}\,\Sigma\mathbf{b}; \tag{1}$$

If we accept Fisher's assumption, the same LDF is obtained in Eq. (2). This equation defines LDF explicitly, whereas Eq. (1) defines LDF implicitly. Therefore, statistical software packages adopt this equation. Some statisticians misunderstand that discriminant analysis is the same as regression analysis. Discriminant analysis is independent of inferential statistics, because there are no SEs of the discriminant coefficients and error rates (Problem 4). Therefore, the leave-one-out (LOO) method [6] was proposed to choose the proper discriminant model.

$$\text{Fisher's LDF} : \mathrm{f}(\mathrm{x}) = {}^{t}\{\mathbf{x} - (\mathbf{m}_1 + \mathbf{m}_2)/2\}\,\Sigma^{-1}(\mathbf{m}_1 - \mathbf{m}_2) \tag{2}$$

Most real data does not satisfy Fisher's assumption. When the variance-covariance matrices of two classes are not the same ($\Sigma_1 \neq \Sigma_2$), the QDF defined in Eq. (3) can be used. The Mahalanobis distance (Eq. (4)) is used for the discrimination of multi-classes, and the Mahalanobis-Taguchi [25] method is applied in quality control.

$$\mathrm{QDF}: \ \mathrm{f}(\mathbf{x}) = {}^{t}\mathbf{x}\left(\Sigma_2^{-1} - \Sigma_1^{-1}\right)\mathbf{x}/2 + \left({}^{t}\mathbf{m}_1\,\Sigma_1^{-1} - {}^{t}\mathbf{m}_2\,\Sigma_2^{-1}\right)\mathbf{x} + \mathrm{c} \tag{3}$$

$$\mathrm{D} = \mathrm{SQRT}\left({}^{t}(\mathbf{x} - \mathbf{m})\Sigma^{-1}(\mathbf{x} - \mathbf{m})\right) \tag{4}$$

These functions are applied in many areas, but cannot be calculated if some features remain constant. There are three cases. First, some features that belong in both classes are the same constant. Second, some features that belong in both classes are different but constant. Third, some feature that belongs to one class is constant. Most statistical software packages exclude all features in these three cases. On the other hand, JMP enhances QDF using the generalized inverse matrix technique. This means that QDF can treat the first and second cases correctly, but cannot handle the third case properly.

Recently, the logistic regression in Eq. (5) has been used instead of LDF and QDF for two reasons. First, it is well known that the error rate of logistic regression is often less than those of LDF and QDF, because it is derived from real data instead of some normal distribution that is liberated from reality. Let 'P' be the probability of belonging to a group of diseases. If the value of some feature is increasing/decreasing, 'P' increases from zero (normal group) to one (group of diseases). This representation is very useful in medical diagnosis, as well as for ratings in real estate and bonds. On the contrary, Fisher's LDF assumes that cases near to the average of the diseases are representative cases of the diseases group. Medical doctors never permit this claim.

$$\mathrm{Log}(\mathrm{P}/(1-\mathrm{P})) = \mathrm{f}(\mathbf{x}) \tag{5}$$

2.2 Before and After SVM

Stam [24] summarized Lp-norm discriminant methods until 1997, and answers the question of "Why have statisticians rarely used Lp-norm methods?" He gives four reasons: Communication, promotion and terminology; Software availability; Relative accuracy of Lp-norm classification methods: Ad hoc studies; and the Accuracy of Lp-norm classification methods: decision theoretic justification. While each of these reasons is true, they are not important. The most important reason is that there is no comparison between these methods with statistical discriminant functions, because discriminant analysis was established by Fisher before mathematical programming (MP) approaches. There are two types of MP applications. The first is modeling by MP, such as for portfolio selection [26], and the second is catch-up modeling, such as for the regression and discriminant analysis. Therefore, the latter type should be compared with preceding results. No statisticians use Lp-norm methods, because there is no research indicating that Lp-norm methods are superior to statistical methods. Liitschwager and Wang [7] defined a model based on the MNM criterion. There are several mistakes, but the most important one is the restriction on the discriminant coefficients. Only one discriminant coefficient should be fixed to −1/1. There is no need to fix the other (k − 1) coefficients in the range [−1, 1].

Vapnik proposed three different SVM models. The hard-margin SVM (H-SVM) indicates the discrimination of linear separable data. H-SVM is defined to maximize the distance of the "Support Vector (SV)" in order to obtain "good generalization" by NLP, which is similar to "not overestimating the validation data in statistics." H-SVM is redefined to minimize (1/"distance of SV") in Eq. (6). This is solved by quadratic programming (QP), which can only be used for linear separable data. This may be why investigation of linear separable data has been ignored. We statisticians misunderstand that discrimination of linear separable data is very easy. In statistics, there was no technical term for linear separable data. However, the condition "MNM = 0" is the same as being linear-separable. Note that "NM = 0" does not imply the data is linear-separable. It is unfortunate that there has been no research into linear separability.

$$\begin{aligned} &\mathrm{MIN} = \|\mathbf{b}\|^2/2; \quad \mathrm{y_i} * ({}^t\mathbf{x_i}\mathbf{b} + \mathrm{b_0}) \geq 1; \quad \mathbf{b} : \mathrm{p-discriminant\ coefficients.} \\ &\mathrm{y_i} = 1/-1 \text{ for } \mathbf{x_i} \in \mathrm{class1/class2.}\ \mathbf{x_i} : \mathrm{p-features(independent\ variables).} \end{aligned} \tag{6}$$

Real data are rarely linear-separable. Therefore, S-SVM has been defined in Eq. (7). S-SVM permits certain cases that are not discriminated by SV ($\mathrm{y_i}*({}^t\mathbf{x_i}\mathbf{b} + \mathrm{b_0}) < 1$).

The second objective is to minimize the summation of distances of misclassified cases (Σe_i) from SV. These two objects are combined by defining some "penalty c." The Markowitz portfolio model to minimize risk and maximize return is the same as S-SVM. However, the return is incorporated as a constraint, and the objective function minimizes risk. The decision maker chooses a solution on the efficient frontier. On the contrary, S-SVM does not have a rule to determine c properly; nevertheless, it can be solved by an optimization solver. (Kernel-SVM is omitted from the research.)

$$\mathrm{MIN} = ||\mathbf{b}||^2/2 + c * \Sigma e_i; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq 1 - e_i; \tag{7}$$

c: penalty c to combine two objectives. e_i: non-negative value.

2.3 IP-OLDF and Revised IP-OLDF

Shinmura and Miyake [12] developed the heuristic algorithm of OLDF based on the MNM criterion. This solves the five features (5-features) model of Cephalo Pelvic Disproportion (CDP) data that consisted of two groups having 19 features. SAS was introduced into Japan in 1978, and three technical reports about the generalized inverse matrix, the sweep operator [5], and SAS regression applications [8] are related to this research. LINDO was introduced to Japan in 1983. Several regression models are formulated by MP [10], e.g., least-squares problems can be solved by QP, and LAV (Least Absolute Value) regression is solved by LP. Without a survey of previous research, the formulation of IP-OLDF can be defined as in Eq. (8). This notation is defined on p-dimensional coefficient space, because the constant is fixed to 1. In pattern recognition, the constant is a free variable. In this case, the model is defined on (p + 1)-coefficient space, and we cannot elicit the same deep knowledge as with IP-OLDF. This difference is very important. IP-OLDF is defined on both p-dimensional data and coefficient spaces. We can understand the relation between the NM and LDF clearly. The linear equation $H_i(\mathbf{b}) = y_i * ({}^t\mathbf{x}_i\mathbf{b} + 1) = 0$ divides p-dimensional space into plus and minus half-planes ($H_i(\mathbf{b}) > 0$, $H_i(\mathbf{b}) < 0$). If $\mathbf{b}_j$ is in the plus half-plane, $f_j(\mathbf{x}) = y_i*({}^t\mathbf{b}_j\mathbf{x} + 1)$ discriminates $\mathbf{x}_i$ correctly, because $f_j(\mathbf{x}_i) = y_i* ({}^t\mathbf{b}_j\mathbf{x}_i + 1) = y_i *({}^t\mathbf{x}_i\mathbf{b}_j + 1) > 0$. On the contrary, if $\mathbf{b}_j$ is included in the minus half-plane, $f_j(\mathbf{x})$ cannot discriminate $\mathbf{x}_i$ correctly, because $f_j(\mathbf{x}_i) = y_i*({}^t\mathbf{b}_j\mathbf{x}_i + 1) = y_i *({}^t\mathbf{x}_i\mathbf{b}_j + 1) < 0$. The n linear equations $H_i(\mathbf{b})$ divide the coefficient space into a finite number of convex polyhedrons. Each interior point of a convex polyhedron has a unique NM that is equal to the number of minus half-planes. We define the OCP as that for which NM is equal to MNM. If $\mathbf{x}_i$ is classified correctly, $e_i = 0$ and $H_i(\mathbf{b}) \geq 0$ in Eq. (8). If there are p cases on $f(\mathbf{x}_i)$ =0, we can obtain the exact MNM. However, if there are over (p+1) cases on $f(\mathbf{x}_i)$=0, this causes the unresolved problem. If $\mathbf{x}_i$ is misclassified, $e_i = 1$ and $H_i(\mathbf{b}) \geq -10000$. This means that IP-OLDF chooses the discriminant hyper-plane $H_i(\mathbf{b}) = 0$ for correctly classified cases, and $H_i(\mathbf{b}) = -10000$ for misclassified cases according to a 0/1 decision variable. if IP–OLDF chooses a vertex having p cases, it chooses the OCP correctly. If it chooses a vertex having over (p+1) cases, it may not choose the OCP. In addition to this defect, IP-OLDF must be solved for the three cases where the constant is equal to 1, 0, −1, because we cannot determine the sign of y_i in advance. Combinations of $y_i = 1/-1$ for $\mathbf{x}_i \in$ class1/class2 are decided by the data, not the analyst.

$$\text{MIN} = \Sigma e_i; \quad H_i(\mathbf{b}) \geq -\ M *\ e_i; \quad M: 10{,}000 \text{ (Big M constant)}. \tag{8}$$

The Revised IP-OLDF in Eq. (9) can find the true MNM, because it can directly find the interior point of the OCP. This means there are no cases where $y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) = 0$. If $\mathbf{x}_i$ is discriminated correctly, $e_i = 0$ and $y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq 1$. If $\mathbf{x}_i$ is misclassified, $e_i = 1$ and $y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) \geq -9999$. It is expected that all misclassified cases will be extracted to alternative SVs, such as $y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) = -9999$. Therefore, the discriminant scores of misclassified cases become large and negative, and there are no cases where $y_i*({}^t\mathbf{x}_i\mathbf{b} + b_0) = 0$. This means that $\mathbf{b}$ is interior point of OCP defined by IP-OLDF.

$$\text{MIN} = \Sigma e_i; \quad y_i * ({}^t\mathbf{x}_i\mathbf{b} +\ b_0) \geq 1 - M * e_i; \quad b_0:\ \text{free decision variable}. \tag{9}$$

If e_i is a non-negative real variable, we utilize Revised LP-OLDF, which is an L1-norm LDF. Its elapsed runtime is faster than that of Revised IP-OLDF. If we choose a large positive number as the penalty c of S-SVM, the result is almost the same as that given by Revised LP-OLDF, because the role of the first term of the objective value in Eq. (7) is ignored. Revised IPLP-OLDF is a combined model of Revised LP-OLDF and Revised IP-OLDF. In the first step, Revised LP-OLDF is applied for all cases, and e_i is fixed to 0 for cases that are discriminated correctly by Revised LP-OLDF. In the second step, Revised IP-OLDF is applied for misclassified cases in the first step. Therefore, Revised IPLP-OLDF can obtain an estimate of MNM faster than Revised IP-OLDF [17, 23], but it is unknown to be free from the unresolved problem.

3 The Unresolved Problem (Problem 1)

3.1 Perception Gap of This Problem

About the unresolved problem, there are several understandings. Most researchers treat the cases $\mathbf{x}_i$ on $f(\mathbf{x}_i) = 0$ in class1. There is no explanation of why it makes sense. Some statisticians explain that it is decided stochastically, because the statistics is a sturdy of probability. This explanation seems theoretically at first glance, but it is nonsense by two reasons. Statistical software adopt the former decision rule because many papers and researchers adopt this rule. In the medical diagnosis, medical doctors strive to judge the patients near by the discriminant hyper-plane. If they know second explanation, they are deeply disappointed in the discriminant analysis. Until now, all LDFs such as Fisher's LDF, logistic regression, H-SVM and S-SVM cannot treat this problem properly. IP-OLDF reveals that only interior points of convex polyhedron can resolve this problem. It can find the vertex of true OCP if data is general position and it stop the optimization choosing p cases on the discriminant hyper-plane. But, it may not find the true MNM if data is not general position and it choose over (p + 1) cases on the discriminant hyper-plane. Revised IP-OLDF can find the interior point of the OCP directly. We cannot judge whether other LDFs choose the interior point, edge or vertex of the convex polyhedron. This is confirmed by checking the number of cases $\mathbf{x}_i$ that satisfy $|f(\mathbf{x}_i)| \leq 10^{-6}$ if we consider $|f(\mathbf{x}_i)| \leq 10^{-6}$ is zero. If this number is zero, this function chooses the interior point of the convex polyhedron. If this number ‘m’ isn’t zero, this LDF chooses the vertex or edge of the convex polyhedron, and true NM has a possibility of increase up to ‘m’

3.2 The Student Data

The student data[1] is proper for us to discuss about the unresolved problem. Fifteen students (y_i = ‘F’) fail the exam and twenty five students (y_i = ‘P’) pass the exam in Table 1. X1 is sturdy hours/day and X2 is expenditure (10,000 yen)/month. X3 is drinking days/week, X4 is sex and X5 is smoking/non-smoking. In the case that IP-OLDF discriminates two classes by (X1, X2), the discriminant hyper-plane of IP-OLDF is X2 = 5. Eight students (X2 > 5) are discriminated to the fail group correctly, four students are on X2 = 5 and three student (X2 < 5) are misclassified into the pass group. On the other hands, twenty one students (X2 < 5) are classified into the pass group correctly and four students are on X2 = 5. Nevertheless IP-OLDF cannot discriminate eight students on X2 = 5, it returns MNM = 3. Revised IP-OLDF can find three discriminant hyper-plane: X2 = 0.006*X1 + 4.984, X2 = 0.25*X1 + 3.65, X2 = 0.99*X1 + 212. And, true MNM = 5. S-SVM (SVM4, c = 10^4) is X2 = X1 + 1, and NM = 6. There is a student having the value of (4, 5) on the discriminant hyper-plane. Therefore, we had better estimated NM = 7. This data is tiny and toy data, but it is useful for the evaluation of the discriminant functions and it is easy for us to understand by scatter plots with two features.

Table 1. The student data.

y_i	F	F	F	F	F	F	F	F	F	F	F	F	F	F	F	P	P	P	P	P
X1	3	1	3	3	2	1	4	3	5	2	3	2	3	3	5	6	9	4	3	2
X2	10	8	7	7	6	6	6	6	5	5	5	5	3	2	2	5	5	5	5	4
y_i	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P	P
X1	5	12	4	10	7	5	7	3	7	7	7	6	3	6	6	8	5	10	9	5
X2	4	4	4	4	4	4	3	3	3	3	3	3	3	3	3	3	3	2	2	2

4 The Discrimination of Linear Separable Data (Problem 2)

4.1 The Importance of This Problem

The purpose of discriminant analysis is to discriminate two classes or objects properly. For this purpose, the discrimination of linear separable data is very important, because we can evaluate the result very clearly. If some LDFs cannot discriminate linear separable data properly, these LDFs should not be used. It is very strange that there is

[1] This data was used for the description of three statistical books using SAS, SPSS and JMP. It is download from (http://sun.econ.seikei.ac.jp/~shinmura/). Click Tab of “Data Archive” and double click “aoyama.xls”.

no research about the discrimination of the linear separable data. H-SVM implies us this discrimination very clearly. But it can be applied only for linear separable data. This may be the reasons why there is no good research about linear separable data until now. Some statistician believes that LDF based on MNM criterion is foolish method, because it over fits for the training samples and its generalization ability may be wrong for the validation samples without examination by real data.

IP-OLDF finds Swiss bank note data is linear separable by two features (X4, X6) and MNMs of 16 models including (X4, X6) are zero. Until now, nobody realize this fact. And, we think it is difficult for us to find linear separable data from real data. But, we can easily obtain two kinds of good research data. First, the pass/fail determination by scores. This is explained in 4.2. Second, every real data is changed to linear separable data by enlarging the distance between the mean of the two classes. The Swiss bank note data consisted of two kinds of bills: 100 genuine and 100 counterfeit bills. There were six features: X1 was the length of the bill (mm); X2 and X3 were the width of the left and right edges (mm), respectively; X4 and X5 were the bottom and top margin widths (mm), respectively; X6 was length of the image diagonal (mm). A total of 63 ($= 2^6-1$) models were investigated. According to Shinmura [18], of the 63 total models, 16 of them including two features (X4, X6) have MNMs of zero; thus, they are linearly separable. The 47 models that remain are not linearly separable. This data is adequate whether or not LDFs can discriminate linearly separable data correctly.

Table 2 shows four results. Upper right (B) is original bank data. Upper left (A) is data expanded to 1.25 times the average distance. Lower left (C) and right (D) are data that are reduced to 0.75 and 0.5 times the average distance. Fisher's LDF is independent of the inferential statistics. But, if we treat $y_i = 1/-1$ as object value and data is analyzed by the regression analysis, obtained regression coefficients are proportional to the discriminant coefficients of Fisher's LDF by the plug-in rule. The stepwise methods can be used formally. 'p' is the number of features by the forward stepwise method. 'Var.' is the selected features. From $p = 1$ to $p = 6$, X6, X4, X5, X3, X2 and X1 are selected in this order by the forward stepwise method. In the regression analysis, Mallow's Cp statistics and AIC are used as variable selection. Usually, the model with minimum of $|Cp - (p + 1)|$ and AIC are recommended. By this rule, Cp statistics choose the same full model. On the other hand, AIC chooses 4-features model (X3, X4, X5, X6) in data 'A'. AIC chooses 5-features model (X2, X3, X4, X5, X6) in other three data.

This table tells us two important facts. We can easily obtained the linear separable data from the real data. The same result as the bank data are observed by the student data, iris data and CPD data those are not linear-separable. Second fact is as follows: "Cp and AIC" choose almost same models, nevertheless 1-feature (X6) model is linear separable in 'A'. And, 2-features (X4, X6) model is linear separable in 'B'. The models selected by "Cp and AIC" are independent from the linear-separablility. Some statisticians don't permit this result by the plug-in rule. On the contrary, they consider Fisher's LDF is the inferential statistics, because it is derived by the Fisher's assumption. This confusion is new problem and is future work (**Future2**).

4.2 Pass/Fail Determination

The pass/fail determination of exam scores makes good research data, because it can be obtained easily, and we can find a trivial discriminant function. My theoretical research starts from 1997 and ends in 2009 [18]. My applied research began in 2010. I negotiated with the National Center for University Entrance Examinations (NCUEE), and borrowed research data consisting of 105 exams in 14 subjects over three years. I finished analyzing the data at the end of 2010, and obtained 630 error rates for Fisher's LDF, QDF, and Revised IP-OLDF. However, NCUEE had requested me not to present the results on March 2011. Therefore, I explain new research results using my statistical exam results. The reason for the special case of QDF and RDA (**Problem 3**) is resolved at the end of 2012. The course consists of one 90 min lecture per week for 15 weeks. In 2011, the course only ran for 11 weeks because of power shortages in Tokyo caused by the Fukushima nuclear accident. Approximately 130 students, mainly freshmen, attended the lectures. Midterm and final exams consisted of 100 questions with 10 choices. Two kinds of pass/fail determinations were discriminated by 100 item scores, and four testlet scores as features. If the pass mark is 50 points, we can easily obtain a trivial discriminant function ($f = T1 + T2 + T3 + T4 – 50$). If $f \geq 0$ or $f < 0$, the student passes or fails the exam, respectively. In this case, students on the discriminant hyper-plane pass the exam, because their score is exactly 50. This indicates that there is no unresolved problem because the discriminant rule is decided by features.

Table 2. Swiss bank data [18].

		A: The distance *1.25				B: Original Bank Data			
Var.	p	Cp	AIC	MNM	LDF	Cp	AIC	MNM	LDF
1-6	6	**7.0**	-863	0	0	**7.0**	-779	0	0
2-6	5	5.3	-865	0	0	5.3	**-781**	0	0
3-6	4	10.5	**-896**	0	0	10.3	-776	0	0
4-6	3	10.9	-859	0	0	10.7	-775	0	0
4, 6	2	118.8	-779	0	0	107.0	-699	0	3
6	1	313.9	-679	0	1	292.0	-604	2	2

		C: The distance * 0.75				D: The distance * 0.5			
Var.	p	Cp	AIC	MNM	LDF	Cp	AIC	MNM	LDF
1-6	6	**7.0**	-676	1	2	**7.0**	-543	5	12
2-6	5	5.3	**-678**	1	2	5.3	**-545**	6	12
3-6	4	9.8	-673	1	1	8.9	-541	7	13
4-6	3	10.1	-673	1	2	8.8	-541	8	14
4, 6	2	97.9	-601	4	6	78.7	-482	16	19
6	1	253.8	-517	6	8	184.4	-417	53	56

4.3 Discrimination by Four Testlets

Table 3 shows the discrimination of four testlet scores as features for 10 % (from third column to seventh column) and 90 % (after eighth column) levels of the midterm exams. The results of 50 % level are omitted. 'p' denotes the number of features selected by the forward stepwise method. In 2010, T4, T2, T1, and T3 are entered in the model selected by the forward stepwise method. The MNM of Revised IP-OLDF and NM of logistic regression are zero in the full model, which means the data is linear-separable in four features. NMs of LDF and QDF are 9 and 2. This means LDF and QDF cannot recognize linear separability. In 2011, Revised IP-OLDF and logistic regression can recognize that the 3-features model (T2, T4, T1) is linear-separable. In 2012, the 2-features model (T4, T2) is linear-separable. T4 and T2 contain easy questions, and T1 and T3 consist of difficult questions for fail group students. This suggests the possibility that pass/fail determination using Revised IP-OLDF can elicit the quality of the test problems and understanding of students in the near future (**Future3**).

Table 3. NMs of four discriminant functions by forward stepwise in midterm exams at the 10 % (from 3rd column to 7th column) and 90 % levels (after 8th column).

	p	Var.	MNM	Logi.	LDF	QDF	Var.	MNM	Logi.	LDF	QDF
2010	1	T4	6	9	11	11	T3	10	37	24	24
	2	T2	2	6	11	9	T4	5	10	20	11
	3	T1	1	3	8	5	T1	**0**	**0**	20	10
	4	T3	**0**	**0**	9	2	T2	0	0	20	11
2011	1	T2	9	17	15	15	T3	6	7	14	14
	2	T4	4	9	11	9	T4	1	1	14	6
	3	T1	**0**	**0**	9	10	T1	**0**	**0**	13	5
	4	T3	0	0	9	11	T2	0	0	14	9
2012	1	T4	4	8	14	12	T3	8	30	12	12
	2	T2	**0**	**0**	11	9	T1	5	12	9	9
	3	T1	0	0	12	8	T4	3	3	10	3
	4	T3	0	0	12	1	T2	**0**	**0**	11	3

Table 4. Summary of error rates of Fisher's LDF and QDF.

		10 %		50 %		90 %	
		LDF	QDF	LDF	QDF	LDF	QDF
Midterm	10	7.5	1.7	2.5	5.0	**16.7**	9.2
	11	7.0	8.5	**2.2**	2.3	10.5	6.7
	12	9.9	**0.8**	4.9	4.8	13.6	7.1
Final	10	4.2	1.7	3.3	4.2	3.3	**10.8**
	11	11.9	2.9	2.9	3.6	3.6	8.6
	12	8.7	2.3	2.3	2.3	13.0	4.5

Table 4 shows a summary of the 18 error rates derived from the NMs of Fisher's LDF and QDF for the linear separable model. Ranges of the 18 error rates of LDF and QDF are [2.2 %, 16.7 %] and [0.8 %, 10.8 %], respectively. Error rates of QDF are lower than those of LDF. At the 10 % level, the six error rates of LDF and QDF lie in the ranges [4.2 %, 11.9 %] and [0.8 %, 8.5 %], respectively. Clearly, the range at the 50 %

level is less than for the 10 % and 90 % levels. Miyake and Shinmura [27] followed Fisher's assumption, and surveyed the relation between population and sample error rates. One of their results suggests that the sample error rates of balanced sample sizes such as 50 % level are close to the population error rates. The above results may confirm this. These results suggest a serious drawback of LDF and QDF based on the variance-covariance matrices. We can no longer trust the error rates of LDF and QDF. Until now, this fact has not been discussed, because there is little research using linear separable data. From this point on, we had best evaluate discriminant functions using linear separable data, because the results are very clear. In genome discrimination, researchers try to estimate the variance-covariance matrices using small sample sizes and large numbers of features. These efforts may be meaningless and lead to incorrect results.

5 Problem 3 (Discrimination of 44 Japanese Cars)

The special cases found in NCUEE exams are confirmed by my exams, also. It is resolved in Nov., 2012. It needs three years because I never doubt the algorithm of QDF and surveyed by the multivariate approach. I checked all features by t-test of two classes, before I abandon the survey. The special case above is more easily explained by the discrimination of 44 Japanese cars.[2] Let us consider the discrimination of 29 regular cars and 15 small cars. Small cars have a special Japanese specification. They are sold as second cars or to women, because they are cost efficient. The emission rate and capacity of small cars are restricted. The emission rate of small and regular cars ranges from [0.657, 0.658] and [0.996, 3.456], respectively. The capacity (number of seats) of small and regular cars are 4 and [5, 8], respectively.

Table 5. Discrimination of Japanese small and regular cars.

p	Var.	t	LDF	QDF[1]	MNM[2]	$\lambda = \gamma = 0.8$	0.5	0.2	0.1
1	Emission	11.37	2	**0**	**0**	2	1	1	**0**
2	Price	5.42	1	**0**	**0**	4	1	0	**0**
3	Capacity	8.93	1	29	**0**	3	1	0	**0**
4	CO_2	4.27	1	29	**0**	4	1	0	**0**
5	Fuel	−4.00	0	29	**0**	5	1	0	**0**
6	Sales	−0.82	0	29	**0**	5	1	0	**0**

[1]If we add small noise to the constant (capacity of small cars), "NMs = 29" are changed to zero.
[2]MNM and NMs of logistic regression are zero.

Table 5 shows the forward stepwise result. At first, "emission" enters the model because the t-value is high. The MNM and NMs of QDF are zero. LDF cannot recognize linear separability. Next, 'price' enters the 2-features model, although the t-value of 'price' is less than that of 'capacity'. In the third step, QDF misclassifies all

[2] This data is open to the paper about DEA (Table 1 in Page 4. http://repository.seikei.ac.jp/dspace/handle/10928/402).

29 regular cars as small cars after “capacity” is included in the 3-features model. This is because the capacity of small cars is fixed to four persons. It is very important that only QDF and RDA are adversely affected by this special case. LDF and the t-test are not affected, because these are computed from the pooled variance of two classes. Modified RDA offers two options such as λ and γ. Four trials show that $\lambda = \gamma = 0.1$ is better than others. JMP division is expected to show the guideline of two options.

6 K-fold Cross Validation (Problem 4)

Usually, the LOO method is used for model selection of the discriminant analysis. In this research, “k-fold cross validation for small sample sizes” is proposed, as it is more powerful than the LOO method. In near future, these results will reveal generalization abilities and the 95 % CIs of Revised IP-OLDF, Revised IPLP-OLDF, Revised LP-OLDF, H-SVM, S-SVM ($c = 10^4$, 1), logistic regression and Fisher’s LDF.

6.1 Hundred-Fold Cross Validation

In the regression analysis, we benefit from inferential statistics, because the SE of regression coefficients, and model selection statistics such as Cp, AIC and BIC, are known a priori. On the other hand, there is no SE of discriminant coefficients and model selection statistics in the discriminant analysis. Therefore, users of the discriminant analysis and SVMs often use the LOO method. Let the sample size be n. One case is used for validation, and the other $(n - 1)$ cases are used as training samples. We evaluate n sets of training and validation samples. If we have a large sample size, we can use k-fold cross validation. The sample is divided into k subsamples. We can evaluate k combinations of the training and validation samples. On the other hand, bootstrap or re-sampling methods can be used with small sample sizes. In this research, large sample sets are generated by re-sampling, and 100-fold cross validation is proposed using these re-sampled data. In this research, “100-fold cross validation for small sample sizes” is applied as follows: (1) We copy 100 times the data from midterm exams in 2012 using JMP. (2) We add a uniform random number as a new variable, sort the data in ascending order, and divide into 100 subsets. (3) We evaluate eight functions such as Revised IP-OLDF, Revised LP-OLDF, Revised IPLP-OLDF, H-SVM, S-SVM ($c = 10^4$ and 1), Fisher’s LDF and logistic regression by 100-fold cross validation using these 100 subsets.

Revised IP-OLDF and S-SVM are analyzed by LINGO [11], developed with the support of LINDO Systems Inc. Logistic regression and LDF are analyzed by JMP, developed with the support of the JMP division of SAS Japan. There is merit in using 100-fold cross validation because we can easily calculate the 95 % CIs of the discriminant coefficients and NMs (or error rates). The LOO method can be used for model selection, but cannot obtain the 95 % CIs. These differences are quite important for analysis of small samples.

6.2 LOO and K–Fold Cross Validation

Table 6 shows the results of the LOO method and NMs in the original data. 'Var.' shows the suffix of four testlet scores named 'T'. Only 11 models were showed, because four 1-feature models were omitted from the table. The MNM of the 2-features model (T2, T4) in No. 6 is zero, as are those of the 4-features model (T1-T4) in No. 1, and the two 3-features models of (T1, T2, T4) in No. 2 and (T2, T3, T4) in No. 3. The NMs of logistic regression and SVM4 ($c = 10^4$) are zero in these four models, but NMs of SVM1 ($c = 1$) are 2 and 3 in No. 2 and No. 6, respectively. It is often observed that S-SVM cannot recognize linear separability when the penalty c has a small value. The LOO method recommends models in No. 3 and No. 6 because these NMs are minimum.

Table 6. LOO and NMs in original test data.

No	Var.	LOO	LDF	Logi	MNM	SVM4	SVM1
1	1–4	14	12	**0**	**0**	**0**	**0**
2	1,2,4	13	12	**0**	**0**	**0**	2
3	2,3,4	**11**	11	**0**	**0**	**0**	**0**
4	1,3,4	15	15	2	2	3	3
5	1,2,3	16	16	6	4	6	6
6	2,4	**11**	11	**0**	**0**	**0**	3
7	1,4	16	16	6	3	6	6
8	3,4	14	13	3	3	4	4
9	1,2	18	17	12	7	7	7
10	2,3	16	11	11	6	11	11
11	1,3	22	21	15	7	10	10

Table 7 shows the results given by Revised IP-OLDF (RIP), SVM4, LDF, and logistic regression (Logi.). The results of SVM1, Revised LP-OLDF and Revised IPLP-OLDF are omitted. First column shows the same No. in Table 6. After four linear separable models, the ranges of seven models are showed. 'MEAN1' column denotes the mean error rate in the training sample. Revised IP-OLDF and logistic regression can recognize linear separability for four models. For SVM4, only model No.1 has an NM of zero. The mean error rates of all Fisher's LDF are over 9.48 %. 'MEAN2' column denotes the mean error rate in the validation sample. Only two models (No.2 and No. 6) of Revised IP-OLDF have NMs of zero and are selected as the best models. The NMs of other functions are greater than zero, and those of LDF are over 9.91 %.

We can conclude that Fisher's LDF is the worst of these four LDFs. Some statisticians believe that NMs of Revised IP-OLDF is less suitable for validation samples, because it over fits for the training samples. On the other hand, Fisher's LDF does not lead to overestimation, because it assumes a normal distribution. These results show that the presumption of 'overestimation' is wrong. We may conclude that real data does not obey Fisher's assumption. To build a theory based on an incorrect assumption will lead to incorrect results [2]. 'Diff.' is the difference between MEAN2 and MEAN1. We think the small absolute value of 'Diff.' implies there is no overestimation. In this sense, Fisher's LDF is better than the other functions, because all values are less than 0.9. However, only high values of the training samples lead to small values of 'Diff.'

'Diff1' denotes the value of (MEAN1 of seven LDFs - MEAN1 of Revised IP-OLDF) in the training samples, and 'Diff2' is the value of (MEAN2 of seven LDFs - MEAN2 of Revised IP-OLDF) in the validation samples. All values of 'Diff1 and Diff2' of SVM4, Fisher's LDF and logistic regression are greater than zero. The maximum values of 'Diff1' given by SVM4, LDF and logistic regression are 2.33, 11.34 and 3.13 %, respectively. And the maximum values of 'Diff2' given by these functions were 1.7, 10.55 and 1.62 %, respectively. It is concluded that Fisher's LDF was not as good as Revised IP-OLDF, S-SVM, and logistic regression by 100-fold cross validation. Therefore, we had better chosen the model of Revised IP-OLDF with minimum value of M2 as the best model. Two models such as (T1, T2, T4) and (T2, T4) are zero. In this case, we had better chosen 2-features model (T2, T4), because of the principle of parsimony or Occam's razor. The values of 'MEAN2' of Revised IP-OLDF, SVM4, Fisher's LDF and logistic regression are 0 %, 1.7 %, 9.91 % and 0.91 %, respectively. This implies that the mean error rates of Fisher's LDF is 9.91 % higher than the best model of Revised IP-OLDF in the validation sample.

Table 7. Comparison of four functions.

RIP	MEAN1	MEAN2	Diff.		
1	**0**	0.07	0.07		
2	**0**	**0**	0		
3	**0**	0.03	0.03		
6	**0**	**0**	0		
4,5,7–11	[0.79,4.94]	[0.03,7.21]	[0.03,2.39]		
SVM4	MEAN1	MEAN2	Diff.	Diff1	Diff2
1	**0**	0.81	0.81	0	0.74
2	0.73	1.62	0.90	0.73	1.62
3	0.13	0.96	0.83	0.13	0.93
6	0.77	**1.70**	0.93	0.77	**1.70**
4,5,7-11	[1.65,6.85]	[3.12,8.02]	[0.66,1.65]	[0.78,2.33]	[0.59,1.36]
LDF	MEAN1	MEAN2	Diff.	Diff1	Diff2
1	9.64	10.54	0.90	9.64	10.47
2	9.89	10.55	0.66	9.89	10.55
3	**9.48**	10.09	0.61	9.48	10.06
6	9.54	**9.91**	0.37	9.54	9.91
4,5,7-11	[10.81,16.28]	[11.03,16.48]	[0.16,0.6]	[7.97,11.34]	[**6.23**,9.61]
Logi	MEAN1	MEAN2	Diff.	Diff1	Diff2
1	**0**	0.77	0.77	0	0.70
2	**0**	1.09	1.09	0	1.09
3	**0**	0.85	0.85	0	0.82
6	**0**	**0.91**	0.91	0	0.91
4,5,7-11	[1.59,7.65]	[2.83,8.04]	[0.35,1.34]	[0.8,3.13]	[0.39,1.62]

In 2014, these results are recalculated using LINGO Ver.14. The elapsed runtimes of Revised IP-OLDF and SVM4 are 3 min 54 s and 2 min 22 s, respectively. The elapsed runtimes of LDF and logistic regression by JMP are 24 min and 21 min, respectively. Reversals of CPU time have occurred for this time.

7 Conclusions

In this research, we have discussed three problems of discriminant analysis. Problem 1 is solved by Revised IP-OLDF, which looks for the interior points of the OCP directly. Problem 2 is theoretically solved by Revised IP-OLDF and H-SVM, but H-SVM can only be applied to linear separable data. Error rates of Fisher's LDF and QDF are very high for linear separable data. This means that these functions should not be used for important discrimination tasks, such as medical diagnosis and genome discrimination. Problem 3 only concerns QDF. This problem was resolved by a t-test after three years of investigation, and can be solved by adding a small noise term to variables. Now, JMP offers a modified RDA, and if we can choose proper parameters, it may be better than LDF and QDF.

However, these conclusions are confirmed by the training samples. In many researches, statistical users have small sample sizes, and cannot evaluate the validation samples. Therefore, "k-fold cross validation for small samples" is proposed. This method confirms the same above conclusion by the validation samples. Many discriminant functions are developed using various criteria after Warmack and Gonzalez [29]. The mission of discrimination should be based on the MNM criterion. Statisticians have tried to develop functions based on the MNM criterion, but this can now be achieved by Revised IP-OLDF using MIP. It is widely believed that Revised IP-OLDF leads to overestimations, but Fisher's LDF is worse for validation samples. Comparison of eight LDFs are examined for future work (**Future4**) by 100-fold cross validation.

References

1. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugenics **7**, 179–188 (1936)
2. Fisher, R.A.: Statistical Methods and Statistical Inference. Hafner Publishing Co., New York (1956)
3. Flury, B., Rieduyl, H.: Multivariate Statistics: A Practical Approach. Cambridge University Press, UK (1988)
4. Friedman, J.H.: Regularized discriminant analysis. J. Am. Stat. Assoc. **84**(405), 165–175 (1989)
5. Goodnight, J.H.: SAS Technical report—the sweep operator: its importance in statistical computing—(R100). SAS Institute Inc., Cary, US (1978)
6. Lachenbruch, P.A., Mickey, M.R.: Estimation of error rates in discriminant analysis. Technometrics **10**, 1–11 (1968)
7. Liitschwager, J.M., Wang, C.: Integer programming solution of a classification problem. Manage. Sci. **24**(14), 1515–1525 (1978)

8. Sall, J.P.: SAS Regression Applications. SAS Institute Inc, Cary (1981). Japanese version is translated by Shinmura, S
9. Sall, J.P., Creighton, L., Lehman, A.: JMP Start Statistics, 3rd edn. SAS Institute Inc, Cary (2004). Japanese version is edited by Shinmura, S
10. Schrage, L.: LINDO—An Optimization Modeling System. The Scientific Press, South San Francisco (1991). Japanese version is translated by Shinmura, S., Takamori, H
11. Schrage, L.: Optimization Modeling with LINGO. LINDO Systems Inc, Chicago (2006). Japanese version is translated by Shinmura, S
12. Shinmura, S., Miyake, A.: Optimal linear discriminant functions and their application. COMPSAC **79**, 167–172 (1979)
13. Shinmura, S.: Optimal linear discriminant functions using mathematical programming. J. Jpn. Soc. Comput. Stat. **11/2**, 89–101 (1998)
14. Shinmura, S.: A new algorithm of the linear discriminant function using integer programming. New Trends. Probab. Stat. **5**, 133–142 (2000)
15. Shinmura, S.: New algorithm of discriminant analysis using integer programming. In: IPSI 2004 Pescara VIP Conference CD-ROM, pp. 1–18 (2004)
16. Shinmura, S.: Overviews of discriminant function by mathematical programming. J. Jpn. Soc. Comput. Stat. **20**(1-2), 59–94 (2007)
17. Shinmura, S.: Improvement of CPU time of revised IPLP-OLDF using linear programming. J. Jpn. Soc. Comput. Stat. **22**(1), 37–57 (2009)
18. Shinmura, S.: The Optimal Linear Discriminant Function. Union of Japanese Scientist and Engineer Publishing, Japanese (2010)
19. Shinmura, S.: Problems of discriminant analysis by mark sense test data. Jpn. Soc. Appl. Stat. **40**(3), 157–172 (2011)
20. Shinmura, S.: Beyond Fisher's Linear Discriminant Analysis—New World of Discriminant Analysis. In: 2011 ISI CD-ROM, pp. 1–6 (2011)
21. Shinmura, S.: Evaluation of Optimal Linear Discriminant Function by 100-fold cross validation. In: 2013 ISI CD-ROM, pp. 1–6 (2013)
22. Shinmura, S.: End of Discriminant Functions Based on Variance Covariance Matrices. In: 2014 ICORE, pp. 5–16 (2014)
23. Shinmura, S.: Improvement of CPU time of linear discriminant functions based on MNM criterion by IP. Stat. Optim. Inf. Comput. **2**, 114–129 (2014)
24. Stam, A.: Non-traditinal approaches to statistical classification: some perspectives on Lp-norm methods. Ann. Oper. Res. **74**, 1–36 (1997)
25. Taguchi, G., Jugulu, R.: The Mahalanobis-Taguchi Strategy—A Pattern Technology System. Wiley, New York (2002)
26. Markowitz, H.M.: Portfolio Selection, Efficient Diversification of Investment. Wiley, New York (1959)
27. Miyake, A., Shinmura, S.: Error rate of linear discriminant function. In: de Dombal, F.T., Gremy, F. (eds.) Decision Making and Medical Care, North-Holland Publishing Company, pp. 435–445. North-Holland Publishing Company, New York (1976)
28. Vapnik, V.: The Nature of Statistical Learning Theory. Springer, Heidelberg (1995)
29. Warmack, R.E., Gonzalez, R.C.: An algorithm for the optimal solution of linear inequalities and its application to pattern recognition. IEEE Trans. Comput. **C-2212**, 1065–1075 (1973)

Springer

http://www.springer.com/978-3-319-17508-9

Operations Research and Enterprise Systems
Third International Conference, ICORES 2014, Angers, France, March 6-8, 2014, Revised Selected Papers
Pinson, E.; Valente, F.; Vitoriano, B. (Eds.)
2015, XII, 301 p. 61 illus., Softcover
ISBN: 978-3-319-17508-9