Improved method for correcting sample Mahalanobis distance without estimating population eigenvalues or eigenvectors of covariance matrix

Yasuyuki Kobayashi¹

Received: 3 June 2019 / Accepted: 23 November 2019 / Published online: 7 December 2019
© The Author(s) 2019

Abstract

The recognition performance of the sample Mahalanobis distance (SMD) deteriorates as the number of learning samples decreases. Therefore, it is important to correct the SMD for a population Mahalanobis distance (PMD) such that it becomes equivalent to the case of infinite learning samples. In order to reduce the computation time and cost for this main purpose, this paper presents a correction method that does not require the estimation of the population eigenvalues or eigenvectors of the covariance matrix. In short, this method only requires the sample eigenvalues of the covariance matrix, number of learning samples, and dimensionality to correct the SMD for the PMD. This method involves the summation of the SMD's principal components (each of which is divided by its expectation obtained using the delta method), Lawley's bias estimation, and the variances of the sample eigenvectors. A numerical experiment demonstrates that this method works well for various cases of learning sample number, dimensionality, population eigenvalues sequence, and non-centrality. The application of this method also shows improved performance of estimating a Gaussian mixture model using the expectation–maximization algorithm.

Keywords Sample Mahalanobis distance · Correction method · Sample eigenvalues and eigenvectors · Delta method · Lawley's bias estimation · Gaussian mixture model

1 Introduction

The Mahalanobis distance (MD) has been used for statistical learning as a basic discriminator under multidimensional normal distributions [1,2]. The MD can be easily defined using the covariance matrix of learning samples and can be calculated faster and more easily than neural networks (NNs) and support vector machines (SVMs). The MD is still widely used for industrial and various other purposes (see a recent example [3]), and there have been cases in which the MD was said to have better recognition performance than NNs and SVMs if there are few learning samples [4–7]. Usually, the population covariance matrix Σ of the learning samples is unknown in advance. Therefore, Hotelling's distance T^2 (sample MD, SMD, or T^2) for a p -variate test vector y is defined as

$$T^2 = (y - \bar{x})' S^{-1} (y - \bar{x}), \quad (1)$$

where the sample covariance matrix S and sample mean vector $\bar{x}$ are calculated from the p -variate learning sample vectors $x_1, \dots, x_n$ that independently follow a p -variate normal distribution $N_p(\mu, \Sigma)$ with population mean μ and population covariance matrix Σ . T^2 follows an F -distribution $F(p, n-p)$ with p and $n-p$ degrees of freedom if y also follows the same distribution $N_p(\mu, \Sigma)$ independently [8]. T^2 follows the F -distribution as in [9]:

$$T^2 = (y - \bar{x})' S^{-1} (y - \bar{x}) \\ \sim \frac{n+1}{n} \frac{p(n-1)}{n-p} F(p, n-p). \quad (2)$$

If μ and Σ are known in advance, the population MD (PMD or D^2) can be defined as

$$D^2 = (y - \mu)' \Sigma^{-1} (y - \mu), \quad (3)$$

and D^2 follows a chi-squared distribution $\chi^2(p)$ with p degrees of freedom if y and $x_1, \dots, x_n$ follow the distribution

✉ Yasuyuki Kobayashi
ykoba@ics.teikyo-u.ac.jp

¹ Faculty of Science and Engineering, Teikyo University,
1-1 Toyosato-dai, Utsunomiya 320-8551, Japan

$N_p(\mu, \Sigma)$ independently [8]. T^2 has a broader distribution than D^2 , and the recognition performance of T^2 is worse than that of D^2 as n decreases and approaches p . This is because the sample eigenvalues of S have a bias from a lack of sufficient learning samples. The large sample eigenvalues tend to be larger than the corresponding population eigenvalues of Σ , whereas the small sample eigenvalues tend to be smaller than the corresponding population eigenvalues. Therefore, some correction must be applied to T^2 in order to improve the recognition performance.

To date, several studies have been conducted to improve T^2 . In order to reduce the bias of the sample eigenvalues of S , $S + \lambda I$ is used with a regularizing constant λ to regularize T^2 [10,11]. For a regularized T^2 , the constant λ is often experimentally determined on an ad hoc basis using a cross-validation method. For example, another method involves determining λ as the average of the $p - m$ smallest sample eigenvalues, where m is determined experimentally [12]. Recently, a kernel using a regularized MD for an SVM was proposed [13]. The distribution of the regularized T^2 obtained from a numerical experiment is unknown, so performing a theoretical analysis of the regularized T^2 is difficult.

Another method involves estimating the unknown Σ using S . The substitution of the estimated population eigenvalues λ_i of Σ for the sample eigenvalues l_i of S in T^2 results in an improvement in the recognition performance of T^2 [14]. Furthermore, the estimated λ_i obtained based on the estimated error of the sample eigenvectors of S using a Monte Carlo simulation provides better performance than when only estimating λ_i [15]. However, the Monte Carlo simulation requires significant computation cost and time. In order to reduce the computation cost and time, the expansion of the factors of the inversed sample eigenvalues, i.e., the eigenvalues of S^{-1} , in T^2 to the second-order Taylor expansion with the variances of the sample eigenvalues obtained from the measurement can be used to correct T^2 [16]. However, there exists a case in which the proposed method provides a corrected S^{-1} that is not positive definite [16]. This method is related to the delta method in statistics [17]. A more precise application of the delta method to S^{-1} results in a formula for which the corrected S^{-1} is always positive definite, and the formula only requires the estimation of the value of λ_i and ratio of $\lambda_i/E[l_i]$ to correct T^2 [18]. However, estimating the unknown λ_i is difficult, and thus, a new correction formula that requires neither the estimation of the value of λ_i nor the ratio of $\lambda_i/E[l_i]$ is proposed [19]. However, the recognition performance of this corrected T^2 is worse than that obtained in [18]. The proposed methods in [18,19] do not take the sample eigenvectors into consideration.

In recent years, a linear metric learning method known as Mahalanobis metric learning has been studied. Mahalanobis metric learning corresponds to a generalization of

the MD and attempts to learn the distances of the form $(x - y)'A(x - y)$, where A is a positive-semidefinite matrix [20]. If A is equal to the inverse of Σ , the distance is equivalent to the MD. Mahalanobis metric learning optimizes A subject to similarity/dissimilarity constraints with regularizers of A (refer to [20]). The regularizers of A are classified into the following three types: Frobenius norm regularization $\|A\|_F^2$ [21–23], linear regularization $\text{tr}(AC)$ [24–27], and information theoretic metric learning $\text{tr}(A) - \log \det A$ [28]. All of these could potentially provide better performance than the MD; however, they require A to be optimized iteratively using a complicated technique, and their distances with the optimized A have an unknown distribution. Therefore, estimating A from only the sample eigenvalues without optimizing the calculation is attractive if it can reduce computation cost and time.

This paper proposes a method for correcting T^2 for D^2 using only the sample eigenvalues of S to define T^2 with dimensionality p and number of the learning samples n . The proposed correction method does not require the population eigenvalues or eigenvectors to be estimated. Therefore, the proposed correction method is apparently easy to use as a discriminator while reducing the computation time and cost. Section 2 discusses the theory for deriving the proposed correction method. Section 3 presents the performance of the proposed correction method compared to that of previous methods using the Monte Carlo simulation. Further, an example of an application for the correction method to improve estimation of a Gaussian mixture model (GMM) by using the expectation–maximization (EM) algorithm is given. Section 4 summarizes the findings and concludes the paper. “Appendix A” lists the common notations used, and “Appendix B” shows the procedures of the simulations in the paper in details.

2 Theory

In this section, Sect. 2.1 introduces some background theories of T^2 and statistical approximation techniques used for correcting T^2 . Section 2.2 derives the proposed correction methods for T^2 considering the effect of both sample eigenvalues and sample eigenvectors of S . Section 2.3 shows previous relative correction methods with only the effect of sample eigenvalues. In order to show the superiority of the proposed correction methods by numerical experiments in Sects. 3, 2.4 summarizes the proposed correction methods and the previous ones.

2.1 Background

Here, let n be the number of learning vector samples $x_1, \dots, x_n$, and let a test vector sample y independently

follow the p -multivariate normal distribution $N_p(\mu, \Sigma)$ with a population mean vector μ and population covariance matrix Σ , where Σ has eigenvalues $\lambda_1 \geq \dots \geq \lambda_p$ and corresponding eigenvectors $\phi_1, \dots, \phi_p$. Then, the learning samples $x_1, \dots, x_n$ give a sample covariance matrix $S = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$, where S has eigenvalues $l_1 \geq \dots \geq l_p$ and the corresponding eigenvectors $f_1, \dots, f_p$.

Definition 1 D^2 is defined by and decomposed with population eigenvalues λ_i and eigenvectors ϕ_i as

$$D^2 = (y - \mu)' \Sigma^{-1} (y - \mu) = \sum_{i=1}^p \frac{\{(y - \mu) \cdot \phi_i\}^2}{\lambda_i} \quad (4)$$

In addition, T^2 is defined by and decomposed with sample eigenvalues l_i and eigenvectors f_i as

$$T^2 = (y - \bar{x})' S^{-1} (y - \bar{x}) = \sum_{i=1}^p \frac{\{(y - \bar{x}) \cdot f_i\}^2}{l_i}. \quad (5)$$

T^2 is different from D^2 because the sample covariance matrix S that is calculated using finite learning samples to define T^2 has an estimation bias from the population covariance matrix Σ . The sample eigenvalues l_i and sample eigenvectors f_i are random variables corresponding to λ_i and ϕ_i ($i = 1, \dots, p$), respectively.

Lemma 1 [29] *The biases for the expectation and variance of l_i , such as*

$$E[l_i] = \lambda_i \left\{ 1 + \frac{1}{n-1} \sum_{j \neq i}^p \frac{\lambda_j}{\lambda_i - \lambda_j} \right\} + O(n^{-2}), \quad (6)$$

$$V[l_i] = \frac{2\lambda_i^2}{n-1} + O(n^{-2}), \quad (7)$$

where it is assumed that $\lambda_1 > \lambda_2 > \dots > \lambda_p$.

Lemma 2 [30] f_i asymptotically follows a normal distribution with λ_i and ϕ_i ($i = 1, \dots, p$) as follows:

$$f_i \sim N_p \left(\phi_i, \frac{\lambda_i}{n-1} \sum_{j \neq i}^p \frac{\lambda_j}{(\lambda_i - \lambda_j)^2} \phi_j \phi_j' \right), \quad (8)$$

where it is assumed that $\lambda_1 > \lambda_2 > \dots > \lambda_p$.

As $n \rightarrow \infty$, l_i converges to λ_i according to (6) and (7) in probability, and f_i also converges to ϕ_i according to (8) in probability. Therefore, T^2 converges to D^2 as $n \rightarrow \infty$. However, it is difficult to calculate the exact distributions of l_i and f_i because their formulae contain zonal polynomials and hypergeometric functions [31]. Therefore, approximated distributions of l_i and f_i are proposed [32,33]. Furthermore,

each principal component term of T^2 , $t_i^2 = \{(y - \bar{x}) \cdot f_i\}^2 / l_i$ in (5), also has a bias depending on the order of magnitude of the sample eigenvalue l_i , which is different from the t -distribution; the distribution of t_i^2 is more concentrated than that of t_j^2 if $i < j$ [34]. However, the exact distribution of t_i^2 is unknown. Therefore, the approximated $E[t_i^2]$ will be considered and applied in Sect. 2.2.

Here, some approximation methods to estimate expectations are shown for use in Sect. 2.2.

Lemma 3 [17] *The expectation of the function $f(X, Y)$ of two random variables X and Y can be approximated using the delta method up to the second order as follows:*

$$\begin{aligned} E[f(X, Y)] &\cong f(E[X], E[Y]) \\ &+ \frac{1}{2} \frac{\partial^2 f}{\partial x^2} \bigg|_{x=E[X], y=E[Y]} V[X] \\ &+ \frac{\partial^2 f}{\partial x \partial y} \bigg|_{x=E[X], y=E[Y]} \text{Cov}[X, Y] \\ &+ \frac{1}{2} \frac{\partial^2 f}{\partial y^2} \bigg|_{x=E[X], y=E[Y]} V[Y], \end{aligned} \quad (9)$$

where $V[X]$ and $V[Y]$ are the variance of X and Y , respectively, and $\text{Cov}[X, Y]$ is the covariance of X and Y . The expectation of $f(X, Y)$ approximated using the delta method up to the zeroth order is

$$E[f(X, Y)] \cong f(E[X], E[Y]). \quad (10)$$

In statistics, the delta method is widely used to approximate the moments of a function of random variables. The delta method is an intuitive technique that involves the use of Taylor's expansion [17].

Lemma 4 [14] $\lambda_i / E[l_i]$ can be approximated by

$$\frac{\lambda_i}{E[l_i]} \cong \frac{n-1}{n+p-2i}. \quad (11)$$

Proof Stein's estimator improves sample eigenvalue l_i of S to l_i^* such that $l_i^* / l_i \cong (n-1)/(n+p-2i)$ [35]. Substituting λ_i for l_i^* and $E[l_i]$ for l_i gives (11). $\square$

Lemma 5 [14] *On the condition that $\lambda_1 \gg \dots \gg \lambda_i \gg \dots \gg \lambda_p$, $\lambda_i / E[l_i]$ can be approximated by*

$$\frac{\lambda_i}{E[l_i]} \cong \frac{n-1}{n-i}. \quad (12)$$

Proof Based on the condition that $\lambda_1 \gg \dots \gg \lambda_i \gg \dots \gg \lambda_p$, $\sum_{j \neq i}^p \lambda_j / (\lambda_i - \lambda_j) \cong 1-i$. Then, Lawley's expectation of l_i (6) is approximated as $E[l_i] \cong \lambda_i \{1 + (1-i)/(n-1)\}$. Therefore, the ratio of $\lambda_i / E[l_i]$ for $\lambda_1 \gg \dots \gg \lambda_p$ is given by (12).

(12) can be also expressed as follows: If $\max\{\lambda_2/\lambda_1, \dots, \lambda_i/\lambda_{i-1}, \dots, \lambda_p/\lambda_{p-1}\} \rightarrow 0$, Ref. [33] shows that

$$l_i \cong \frac{\lambda_i}{n-1} \chi^2(n-i). \quad (13)$$

(12) can also be obtained from the expectation of (13). Furthermore, (7) can be obtained from the variance of (13) if $n \gg i$. $\square$

2.2 Proposed models

In order to correct the bias of T^2 more precisely, each t_i^2 should be corrected according to its own bias.

Lemma 6 *The bias of T^2 can be corrected by dividing each t_i^2 by $E[t_i^2]$, i.e., the expectation of t_i^2 :*

$$CT^2 = \sum_{i=1}^p \left(\frac{t_i}{\sqrt{E[t_i^2]}} \right)^2, \quad (14)$$

where $t_i = (y - \bar{x}) \cdot f_i / \sqrt{l_i}$ is the studentized principal component corresponding to l_i in T^2 .

The expectation of t_i^2 divided by $E[t_i^2]$ is therefore one by the analogy that the expectation of d_i^2 is equal to one, where $d_i^2 = \{(y - \mu) \cdot \phi_i\}^2 / \lambda_i$, i.e., each principal component term of D^2 in (4) follows $\chi^2(1)$, a chi-squared distribution with one degree of freedom. However, it is difficult to calculate the exact distributions of l_i and f_i , and it is necessary to approximate $E[t_i^2]$ precisely. In order to apply the delta method for obtaining $E[t_i^2]$, let us assume that $y - \bar{x}$, l_i , and f_i are random variables and ϕ_i and λ_i are not.

Theorem 1 *If f_i is given by a linear combination of ϕ_i , using the delta method approximates t_i^2 as*

$$t_i^2 = \frac{1}{C_i l_i} \left(\lambda_i z_i^2 + \sum_{\substack{j=i \\ k \neq i}}^p \sqrt{\lambda_i \lambda_k} z_i z_k a_{ik} + \sum_{\substack{j \neq i \\ k=i}}^p \sqrt{\lambda_i \lambda_j} z_i z_j a_{ij} \right. \\ \left. + \sum_{\substack{j \neq i \\ k \neq i, j \neq k}}^p \sqrt{\lambda_j \lambda_k} z_j z_k a_{ij} a_{ik} + \sum_{\substack{j \neq i \\ j=k}}^p \lambda_j z_i^2 a_{ij}^2 \right), \quad (15)$$

where a_{ij} denotes the coefficients of the linear combination of ϕ_i , C_i is the function of the squared a_{ij} , and z_i is a random variable that follows $N(0, 1)$ for $i = 1, \dots, p$, $j \neq i$.

Proof As n is generally large, let us suppose $y - \bar{x}$ follows the distribution $N_p(\mu, \Sigma)$; $y - \bar{x}$ is then given by

$$y - \bar{x} = \sum_{i=1}^p z_i \sqrt{\lambda_i} \phi_i, \quad (16)$$

On applying (16) to t_i^2 , we obtain

$$t_i^2 = \frac{(y - \bar{x})' f_i f_i' (y - \bar{x})}{l_i} \\ = \sum_{j,k=1}^p \frac{z_j z_k \sqrt{\lambda_j \lambda_k} \phi_j' f_i f_i' \phi_k}{l_i}. \quad (17)$$

According to (8), f_i can be decomposed into a linear combination of ϕ_j ($j = 1, \dots, p$) with sufficiently small random coefficients a_{ij} , i.e.,

$$f_i = \phi_i + \sum_{j \neq i}^p a_{ij} \phi_j, \quad (18)$$

$$a_{ij} \sim N\left(0, \frac{\lambda_i \lambda_j}{(n-1)(\lambda_i - \lambda_j)^2}\right) \quad (j \neq i). \quad (19)$$

In order to satisfy the condition that the sample eigenvector f_i is a unit vector even if a_{ij} is large, the normalized f_i in (18) can be written as

$$f_i = \frac{\phi_i + \sum_{j \neq i}^p a_{ij} \phi_j}{\sqrt{C_i}}, \quad (20)$$

$$C_i = 1 + \sum_{k \neq i}^p a_{ik}^2. \quad (21)$$

As $\phi_i' \phi_j = \delta_{ij}$, i.e., the Kronecker delta, on applying (20) to $\phi_j' f_i f_i' \phi_k$ in (17), we obtain

$$\phi_j' f_i f_i' \phi_k = \begin{cases} 1/C_i & (i = j = k), \\ a_{ik}/C_i & (i = j \neq k), \\ a_{ij}/C_i & (j \neq i = k), \\ a_{ij} a_{ik}/C_i & (i \neq j \neq k), \\ a_{ij}^2/C_i & (i \neq j = k). \end{cases} \quad (22)$$

On applying (22) to (17), we obtain (15), where all the random variables C_i , l_i , z_i , and a_{ij} ($i = 1, \dots, p$, $j \neq i$) in (15) are independent, and thus, the covariance of each pair of random variables is zero and does not appear in the resultant formulae of the delta method obtained from (15). $\square$

Theorem 2 If λ_i and $E[l_i]$ are known in advance, using the delta method we can approximate $E[t_i^2]$ of (15) as

$$E[t_i^2] \cong \frac{\lambda_i}{E[l_i]} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_i}{E[l_i]} \right)^2 \right\} \cdot \frac{1 + \sum_{j \neq i}^p \frac{\lambda_j^2}{(n-1)(\lambda_i - \lambda_j)^2}}{1 + \sum_{j \neq i}^p \frac{\lambda_i \lambda_j}{(n-1)(\lambda_i - \lambda_j)^2}}. \quad (23)$$

Proof Let us suppose that the delta method is applied up to the second order for l_i by using $V[l_i] = 2\lambda_i^2/(n-1)$ in (7) and $E[l_i]$, up to the zeroth order for a_{ij} by using $E[a_{ij}] = 0$, up to the zeroth order for a_{ij}^2 by using $E[a_{ij}^2] = \lambda_i \lambda_j / \{(n-1)(\lambda_i - \lambda_j)^2\}$, up to the zeroth order for C_i by using $E[C_i] = 1 + \sum_{j \neq i}^p E[a_{ij}^2]$, up to the zeroth order for z_i^2 by using $E[z_i^2] = 1$, up to the zeroth order for $z_i z_j$ by using $E[z_i z_j] = 0$, and up to the zeroth order for $a_{ij} a_{ik}$ by using $E[a_{ij} a_{ik}] = 0$. $1/l_i$ in (15) is applied with the delta method up to the second order, and $E[1/l_i]$ is given by

$$E\left[\frac{1}{l_i}\right] \cong \frac{1}{E[l_i]} + \frac{V[l_i]}{E[l_i]^3} = \frac{1}{E[l_i]} + \frac{2\lambda_i^2}{n-1} \frac{1}{E[l_i]^3}, \quad (24)$$

$1/C_i$ in (15) is applied with the delta method up to the zeroth order, and $E[1/C_i] \cong 1/E[C_i]$. $\square$

Thus, T^2 is approximately corrected by using (14) with (23). In the case of (23), if $n \rightarrow \infty$, then $E[l_i] \rightarrow \lambda_i$, and thus $E[t_i^2] \rightarrow 1$ in (23). Moreover, $t_i^2 \rightarrow d_i^2 = \{(y - \mu) \cdot \phi_i\}^2 / \lambda_i$. Therefore, correcting T^2 by using (14) with (23) has no effect if $n \rightarrow \infty$. The first term in (23), i.e.,

$$\frac{\lambda_i}{E[l_i]} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_i}{E[l_i]} \right)^2 \right\} \quad (25)$$

indicates the effect of the bias and variance of l_i in the denominator in (5). This shows that substituting λ_i for l_i in (5) is insufficient for correcting T^2 . Furthermore, the second term in (23), i.e.,

$$\frac{1 + \sum_{j \neq i}^p \frac{\lambda_j^2}{(n-1)(\lambda_i - \lambda_j)^2}}{1 + \sum_{j \neq i}^p \frac{\lambda_i \lambda_j}{(n-1)(\lambda_i - \lambda_j)^2}} \quad (26)$$

indicates the effect of the variances of f_i , and the variance becomes large if the neighbors of λ_i approach λ_i . However, if $\lambda_i \cong \lambda_j$ for only one pair of λ_i and λ_j , (26) approaches λ_j / λ_i and $E[T^2]$ of (23) is given by the principal components for $\lambda_i \cong \lambda_j$, i.e.,

$$E[T^2] \cong \frac{\lambda_j}{E[l_i]} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_i}{E[l_i]} \right)^2 \right\} + \frac{\lambda_i}{E[l_j]} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_j}{E[l_j]} \right)^2 \right\}. \quad (27)$$

Therefore, (23) reflects all the effects of l_i and f_i in terms of all λ_j ($j = 1, \dots, p$).

However, if there is only a restricted number of unknown learning samples, it is difficult to estimate λ_i and the true $E[l_i]$ precisely and to apply (23) to (14) for correcting T^2 . Hence, it would then be required to estimate λ_i and the true $E[l_i]$ approximately, with as low a cost as possible. Then, approximating $\lambda_i / E[l_i]$ as (11) or (12) can satisfy the above condition.

Theorem 3 The approximated $E[t_i^2]$ including the effect of the variance of sample eigenvectors f_i are given by

$$E[t_i^2] \cong \frac{n-1}{n+p-2i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n+p-2i} \right)^2 \right\} \cdot \frac{1 + \sum_{j \neq i}^p \frac{l_j^2}{(n-1)(l_i - l_j)^2}}{1 + \sum_{j \neq i}^p \frac{l_i l_j}{(n-1)(l_i - l_j)^2}}, \quad (28)$$

or

$$E[t_i^2] \cong \frac{n-1}{n-i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n-i} \right)^2 \right\} \cdot \frac{1 + \sum_{j \neq i}^p \frac{l_j^2}{(n-1)(l_i - l_j)^2}}{1 + \sum_{j \neq i}^p \frac{l_i l_j}{(n-1)(l_i - l_j)^2}}. \quad (29)$$

Both (28) and (29) can be calculated without estimating λ_i or $E[l_i]$.

Proof On substituting l_i for λ_i and applying (11) or (12) to $\lambda_i / E[l_i]$ in (23), we obtain (28) and (29). $\square$

Thus, T^2 can be corrected without estimating λ_i or ϕ_i by applying $E[t_i^2]$ from (28) or (29) to (14).

It is necessary to compare the performance of (28) and (29) with that of models without the effect of the variance of f_i .

Corollary 1 An approximated $E[t_i^2]$ not including the effect of the variance of f_i is given by

$$E[t_i^2] \cong \frac{n-1}{n-i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n-i} \right)^2 \right\}. \quad (30)$$

Proof (30) is given by omitting (26) from (29). (30) includes no correction from the effects of f_i , as it is assumed that $f_i \cong \phi_i$. $\square$

Table 1 List of all the Mahalanobis distances considered in the numerical experiments

Abbr.	Formulae of studentized principal component score and MD	Eqn. No.
ppl.	$z_i = \frac{(y - \mu) \cdot \phi_i}{\sqrt{\lambda_i}}, D^2 = \sum_{i=1}^p z_i^2.$	(33)
smd.	$t_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{l_i}}, T^2 = \sum_{i=1}^p t_i^2.$	(34)
pd	$ct_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{\lambda_i}} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_i}{E[l_i]} \right)^2 \right\}^{-\frac{1}{2}}, CT^2 = \sum_{i=1}^p ct_i^2.$	(35)
Sd	$ct_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{l_i}} \left[\frac{n-1}{n+p-2i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n+p-2i} \right)^2 \right\} \right]^{-\frac{1}{2}}, CT^2 = \sum_{i=1}^p ct_i^2.$	(36)
fSd	$ct_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{l_i}} \left[\frac{n-1}{n+p-2i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n+p-2i} \right)^2 \right\} \frac{1 + \sum_{j \neq i}^p \frac{l_j^2}{(n-1)(l_i - l_j)^2}}{1 + \sum_{j \neq i}^p \frac{l_i l_j}{(n-1)(l_i - l_j)^2}} \right]^{-\frac{1}{2}}, CT^2 = \sum_{i=1}^p ct_i^2.$	(37)
Ld	$ct_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{l_i}} \left[\frac{n-1}{n-i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n-i} \right)^2 \right\} \right]^{-\frac{1}{2}}, CT^2 = \sum_{i=1}^p ct_i^2.$	(38)
fLd	$ct_i = \frac{(y - \bar{x}) \cdot f_i}{\sqrt{l_i}} \left[\frac{n-1}{n-i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n-i} \right)^2 \right\} \frac{1 + \sum_{j \neq i}^p \frac{l_j^2}{(n-1)(l_i - l_j)^2}}{1 + \sum_{j \neq i}^p \frac{l_i l_j}{(n-1)(l_i - l_j)^2}} \right]^{-\frac{1}{2}}, CT^2 = \sum_{i=1}^p ct_i^2.$	(39)

Applying (28) and (29) to (14) can correct the bias of T^2 (5) from the randomness of l_i and f_i for D^2 (4). However, (30) can correct the bias of T^2 (5) from the randomness of l_i .

2.3 Related models as $f_i \cong \phi_i$ in previous studies

Corollary 2 [18] Assuming that the true values of λ_i and $E[l_i]$ are given, another approximated $E[t_i^2]$ assuming $f_i \cong \phi_i$ (ignoring the effect of the variance of f_i) is

$$E[t_i^2] \cong \frac{\lambda_i}{E[l_i]} \left\{ 1 + \frac{2}{n-1} \left(\frac{\lambda_i}{E[l_i]} \right)^2 \right\}, \quad (31)$$

which is the first factor on the right side of (23).

However, it is unusual for the true values of λ_i to be known in advance.

Corollary 3 [19] Without estimating λ_i , by applying (11) to (31), another approximated $E[t_i^2]$ assuming $f_i \cong \phi_i$ is as follows:

$$E[t_i^2] \cong \frac{n-1}{n+p-2i} \left\{ 1 + \frac{2}{n-1} \left(\frac{n-1}{n+p-2i} \right)^2 \right\}. \quad (32)$$

On applying (32) to (14), T^2 is corrected well such that the recognition performance of the corrected T^2 is improved from that of the uncorrected T^2 . However, a more accurate

correction of T^2 can be obtained using (31) than that obtained using (32) if the true values of λ_i are given in advance. It has been reported that applying the delta method for t_i^2 up to the third or higher order does not improve the recognition performance of the corrected T^2 [19].

2.4 Summary of corrected T^2 , etc

Table 1 summarizes the formulae of D^2 , uncorrected T^2 , and the corrected T^2 (CT^2) that are proposed above. The abbreviations in Table 1 are also shown to indicate the corresponding D^2 or T^2 in Sect. 3, and each ct_i in Table 1 is the corrected studentized principal component score, i.e., the principal component score divided by $\sqrt{l_i}$ applied with the proposed $E[t_i^2]$ in the CT^2 .

Table 1 indicates each method as follows:

- “ppl.” (33) corresponds to D^2 (4) in Definition 1, which the CT^2 should aim to follow.
 - “smp.” (34) corresponds to T^2 (5) in Definition 1, which is not corrected.
 - “pd” (35) corresponds to the CT^2 using (31) in Corollary 2, assuming that the true values of λ_i and $E[l_i]$ are given [18].
- The following methods from “Sd” to “fLd” assume that λ_i and $E[l_i]$ are unknown.
- “Sd” (36) corresponds to the CT^2 using (32) in Corollary 3, assuming that $f_i \cong \phi_i$ [19].

The following methods from “fSd” to “fLd” are proposed in this paper for the first time.

- “fSd” (37) corresponds to the CT^2 using (28) in Theorem 3, assuming that $f_i \neq \phi_i$.
- “Ld” (38) corresponds to the CT^2 using (30) in Corollary 1, assuming that $f_i \cong \phi_i$.
- “fLd” (39) corresponds to the CT^2 using (29) in Theorem 3, assuming that $f_i \neq \phi_i$.

3 Numerical experiments

3.1 Monte Carlo simulation of MDs

In the numerical experiments, the dependence of the corrected MDs on the non-centrality δ , number of learning samples n , and dimensionality p is considered using Monte Carlo simulation. The detailed procedure of the Monte Carlo simulation is shown in “Appendix B.1”. Furthermore, the dependence on the population eigenvalues λ_i should be considered because the correcting methods of fSd (37), Ld (38), and fLd (39) derived from (6) and (8) are based on the assumption that $\lambda_1 > \lambda_2 > \dots > \lambda_p$. Thus, two sequence types of $\lambda_i (i = 1, \dots, p)$ are considered. Sequence 1 is a geometric sequence between $\lambda_1 = 1$ and $\lambda_p = 10^{-10}$ for $\lambda_1 > \lambda_2 > \dots > \lambda_p$. Sequence 2 is a geometric sequence between $\lambda_1 = 1$ and $\lambda_p = 1$ for $\lambda_1 = \lambda_2 = \dots = \lambda_p$. Eqn. (26) does not hold for Sequence 2. However, (26) holds with the substitution of l_i for λ_i because the sample eigenvalue l_i has a bias from λ_i and no two values of l_i are equal; therefore, fSd (37), Ld (38), and fLd (39) can hold even if $\lambda_1 = \lambda_2 = \dots = \lambda_p$.

In order to eliminate as much of the sample error as possible from the empirical PDCs, n_{Σ} is fixed as 10,000 and n_y is fixed as 1,000. Therefore, the number of test samples is equal to $n_{\Sigma} \cdot n_y = 10,000,000$. The MDs considered are listed in Table 1: D^2 (ppl.), T^2 (smp.), and five of the corrected MDs (pd, Sd, fSd, Ld, and fLd).

3.1.1 Results of sequence 1 ($\lambda_1 > \dots > \lambda_p$)

In order to compare the correcting performance of T^2 for D^2 , the Hellinger distance (44) of the MDs from D^2 (ppl.) is one of the most accessible metrics. Figures 1, 2 and 3 show the dependence of the Hellinger distances from D^2 on the non-centrality δ , number of learning samples n , and dimensionality p , respectively. The conditions for δ , n , and p in each case are written in the respective figures. Figure 1 shows that, as the non-centrality δ increases, all the distances also increase gradually. Figure 2 shows that, as n increases, all the distances decrease rapidly. Figure 3 shows that, as p increases, all the distances also increase gradually. Overall, all the corrected MDs are closer to D^2 than T^2 . In particular,

Fig. 1 Dependence of the Hellinger distances from D^2 (ppl.) on the non-centrality δ for Sequence 1

Fig. 2 Dependence of the Hellinger distances from D^2 (ppl.) on the number of learning samples n for Sequence 1

Fig. 3 Dependence of the Hellinger distances from D^2 (ppl.) on the dimensionality p for Sequence 1

the models fLd (39) and pd (35) are the closest to D^2 of all the MDs from Figs. 1, 2 and 3. However, the model pd (35) requires the population eigenvalue λ_i in advance. Therefore, the model fLd (39) is the best corrected MD if the population eigenvalue λ_i is unknown in advance.

Here, we confirm that the model fLd (39) is the best corrected MD if λ_i is unknown in advance. Figure 4 shows the PDCs of the MDs. The PDCs of fLd (39) and pd (35) are the closest to that of ppl. (D^2) (33). Figures 5 and 6 show

Fig. 4 Probability density of normalized MDs for Sequence 1

Fig. 5 Dependence of expectations on the non-centrality δ for Sequence 1

Fig. 6 Dependence of variances on the non-centrality δ for Sequence 1

the dependences of the expectation and the variance on the non-centrality δ . Both of fLd (39) and pd (35) are the closest to that of ppl. (33). However, the variances of both fLd (39) and pd (35) are still larger than that of ppl. (33). This implies that neither the fLd (39) nor pd (35) model can perfectly correct the right long tails of the PDCs, such as that of ppl. (33), as seen in Fig. 4. Figure 7 shows the misrecognition rates $(\alpha + \beta)_{\min}$ of the MDs. $(\alpha + \beta)_{\min}$ of pd (35) is the second smallest of all, and that of fLd (39) is the third smallest. However, the difference between the two is small in Fig. 7. Figure 8 shows the threshold Θ at $(\alpha + \beta)_{\min}$ of the MDs. The thresholds of ppl. (33), fLd (39), and pd (35) are almost the same and the smallest in Fig. 8. Therefore, the model

Fig. 7 Dependence of $(\alpha + \beta)_{\min}$ on the non-centrality δ for Sequence 1

Fig. 8 Dependence of threshold Θ at $(\alpha + \beta)_{\min}$ on the non-centrality δ for Sequence 1

fLd (39) can correct T^2 for D^2 to improve the recognition performance of the MDs if λ_i or ϕ_i is unknown in advance.

3.1.2 Results of sequence 2 ($\lambda_1 = \dots = \lambda_p$)

As mentioned in the preface in Sect. 3.2, an investigation of the case in which the assumption that $\lambda_1 > \lambda_2 > \dots > \lambda_p$ does not hold is required. Figures 9, 10 and 11 show the dependence of the Hellinger distances from D^2 on the non-centrality δ , number of learning samples n , and dimensionality p , respectively. Figure 9 shows that as the non-centrality δ increases, all the distances also increase gradually. However, the Hellinger distance of pd (35) in Fig. 9 is larger than that of fLd (39) in Sequence 2, in contrast to Fig. 1 for Sequence 1. This is because the PDC of pd (35) is overbiased to the left compared to that of ppl. (33) in Fig. 12. Figure 13 shows that the misrecognition rate $(\alpha + \beta)_{\min}$ of pd (35) is closer to that of ppl. (33) than that of fLd (39). However, Fig. 14 shows that the threshold Θ of pd (35) is smaller than that of ppl. (33) because pd (35) is overbiased in Fig. 12. The relation of the Hellinger distance between pd (35) and fLd (39) becomes reversed as n increases in Fig. 10 or p increases in Fig. 11. This is because the PDC of pd (35) becomes closer to that of ppl. (33) as n increases or p

Fig. 9 Dependence of the Hellinger distances from D^2 (ppl.) on the non-centrality δ for Sequence 2

Fig. 10 Dependence of the Hellinger distances from D^2 (ppl.) on the number of learning samples n for Sequence 2

Fig. 11 Dependence of the Hellinger distances from D^2 (ppl.) on the dimensionality p for Sequence 2

increases (graphs of the PDCs are not shown in this paper). However, pd (35) is difficult to calculate if λ_i is unknown in advance. Therefore, the model fLd (39) can also correct T^2 for D^2 to improve the recognition performance of MDs if λ_i or ϕ_i is unknown in advance even in Sequence 2.

From all the results of the numerical experiments, it is observed that the model fLd (39) can best correct T^2 for D^2 without estimating λ_i or ϕ_i with respect to the dependence on the non-centrality δ , number of learning samples n , and

Fig. 12 Probability density of normalized MDs for Sequence 2

Fig. 13 Dependence of $(\alpha + \beta)_{\min}$ on the non-centrality δ for Sequence 2

Fig. 14 Dependence of threshold Θ at $(\alpha + \beta)_{\min}$ on the non-centrality δ for Sequence 2

dimensionality p for both $\lambda_1 > \dots > \lambda_p$ and $\lambda_1 = \dots = \lambda_p$. This is because fLd (39) corrects both the sample eigenvalue l_i in the denominator and sample eigenvector f_i in the numerator of the decomposed principal component score of T^2 . It is interesting that Lawley's approximation (12) appears to be better than Stain's approximation (11) for the estimation of $\lambda_i/E[l_i]$, even if $\lambda_1 = \dots = \lambda_p$. However, the reason for this is still an open question.

Table 2 Results of applying corrected SMD to estimate a GMM by using the EM algorithm

Type of MD	L.-L. of $p(x \theta)$	Classification rate		
		Class 1 (%)	Class 2 (%)	Class 3 (%)
smd. (34)	−16.38	95	77	85
fLd (39)	−15.05	100	66	85

3.2 Example of application

In order to evaluate the performance of correcting T^2 for D^2 for an application, the model fLd (39) is applied to the EM algorithm for GMMs. A p -variate GMM with k components has the density as the sum of probability density functions of normal distribution $N_p(m, S)$:

$$p(x|\theta) = \sum_{j=1}^k w_j g(x|m_j, S_j), \quad (40)$$

$$g(x|m_j, S_j) = \frac{1}{\sqrt{(2\pi)^p |S_j|}} e^{-\frac{1}{2}(x-m_j)' S_j^{-1} (x-m_j)}, \quad (41)$$

where $\theta = \{(w_j, m_j, S_j)\}_{j=1}^k$, and w_j is the probability that the observed sample x is generated by the j th component and satisfies $w_j > 0$ and $\sum_{j=1}^k w_j = 1$. The EM algorithm for a GMM [36] with k components given n independent and identically distributed samples $x_1, x_2, \dots, x_n \in R^p$ is shown in “Appendix B.2”.

After running the EM algorithm, a sample y is assigned to the component j for which

$$\left\{ (y - m_j)' S_j^{-1} (y - m_j) + \log |S_j| - 2 \log w_j \right\} / p \quad (42)$$

is the smallest. $(x - m_j)' S_j^{-1} (x - m_j)$ in (41) for a GMM, and (42) is a type of T^2 , and n_j , which is the effective number of the learning samples of the component j , holds $n_j \leq n$. We contend that the well-known EM algorithm for a GMM (see [36]) does not consider that S_j is different from Σ_j if n_j is not sufficiently large. Therefore, $g(x|m_j, S_j)$ tends to have a bias from a lack of sufficient learning samples, and the estimations of the EM algorithm become worse unless $p \ll n_j$. In order to reduce the bias, the application of fLd (39) to $(x - m_j)' S_j^{-1} (x - m_j)$ in (41) and (42) is investigated.

The dataset “Wine recognition data” in Ref. [37] was used for classifying by means of the EM algorithm for a GMM. These data are the results of a chemical analysis of wines produced in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents ($p = 13$) found in each of the three types of wines ($k = 3$). The numbers of instances are $n_1 = 59$ for class 1, $n_2 = 71$ for class 2, and $n_3 = 48$ for class 3 (total $n = 178$). As each class does not satisfy $n_j \gg p$, this dataset appears to have a bias from a lack of sufficient learning sam-

ples. After running the EM algorithm, the sample eigenvalues l_i of each component range from 10^4 to 10^{-3} , which corresponds to Sequence 1 in Sect. 3.1.1. This suggests that fLd (39) is appropriate for the dataset.

Table 2 summarizes the results. For log-likelihood (L.-L.) of $p(x|\theta)$ (40), the L.-L. of fLd (39) is larger than that of smd. (34). This means that by using fLd (39) the GMM can be estimated more correctly than by using smd. (34), i.e., the conventional SMD used in the EM algorithm for a GMM. Thus, fLd (39) improves the classification rate of Class 1 in comparison to smd. (34). However, smd. (34) works better than fLd (39) for the classification rate of Class 2. This is because the EM algorithm refines the estimates of the whole model of a GMM to improve its likelihood, not that of each component. Another method should be considered to improve the classification rate of each component.

Therefore, correcting T^2 for D^2 with fLd (39) can improve the estimation performance for models including formulae like T^2 , such as a GMM, by simply replacing T^2 with fLd (39) without estimating the true eigenvalues or eigenvectors of the covariance matrix in T^2 in advance. The above GMM example does not have singular sample covariance matrices. Therefore, the model fLd (39) can be applied to improve estimation performance. However, the EM algorithm for GMMs sometimes has a singularity problem; S_j of a fitted component j has very small positive eigenvalues when the data of S_j are in a low-dimensional subspace [36,38], in which case the model fLd (39) cannot correct T^2 . One of the methods to avoid the problem is regularizing S_j to be a low-rank matrix [39]. Hence, the model fLd (39) should be extended for a low-rank S_j in the future.

4 Conclusion

In order to correct the SMD or T^2 for PMD or D^2 , this paper presents a correction method (called fLd in this paper) that requires only the sample eigenvalues of the sample covariance matrix, dimensionality p , and number of learning samples n . The proposed method requires neither the estimation of the population eigenvalues nor eigenvectors nor the use of optimization or time-consuming methods such as the Monte Carlo method. The proposed method eliminates the bias of each principal component score of T^2 by dividing the principal component score by its expectation. By using

the delta method in statistics, Lawley's bias estimation of the sample eigenvalues, and the variances of the sample eigenvectors, the approximated expectation can be obtained by using the sample eigenvalues, p , and n . A numerical experiment using the Monte Carlo method to generate test samples was conducted to investigate the correction performance of the proposed method under various conditions, such as non-centrality δ , n , p , and two types of population eigenvalue sequences: $\lambda_1 > \lambda_2 > \dots > \lambda_p$ and $\lambda_1 = \lambda_2 = \dots = \lambda_p$. Although Lawley's bias estimation of the sample eigenvalues is based on the assumption that $\lambda_1 > \lambda_2 > \dots > \lambda_p$, the proposed method substitutes the sample eigenvalues for the population eigenvalues in the formula and corrects T^2 well even if $\lambda_1 = \lambda_2 = \dots = \lambda_p$. In addition, the dependence of the corrected or uncorrected T^2 on δ shows that the recognition performance, i.e., the minimum misrecognition rate of T^2 corrected using the proposed method is improved from that of the uncorrected T^2 . Furthermore, the application of the proposed method shows an improvement in estimating a GMM by using the EM algorithm. Thus, it is expected that the proposed model can be applied to other models including terms such as the SMD. However, the proposed method still has room to correct T^2 for D^2 perfectly. The PDC of the corrected T^2 is closer to that of D^2 than any other previously proposed corrected T^2 , although it does not coincide with that of D^2 . The minimum misrecognition rate of the corrected T^2 decreases but is not equal to that of D^2 . Moreover, the probability distribution of the corrected T^2 is currently theoretically unknown. The above result suggests that there is room for improvement in the elimination of the bias of T^2 using the proposed method. In order to improve the T^2 correction, some new ideas should be considered, such as theoretically elaborating the probability density function of each studentized principal component score of T^2 , improving Lawley's bias approximation, and the modeling of sample eigenvectors.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number JP15H02798.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your

intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

A Common notations

p	Dimensionality
n	Number of learning samples
μ	Population mean vector
Σ	Population covariance matrix
$\lambda_1, \dots, \lambda_p$	Population eigenvalues of Σ
$\phi_1, \dots, \phi_p$	Population eigenvectors of Σ
$x_1, \dots, x_n$	Learning sample vectors
y	Test sample vector
$\bar{x}$	Sample mean vector of $x_1, \dots, x_n$
S	Sample covariance matrix of $x_1, \dots, x_n$
$l_1, \dots, l_p$	Sample eigenvalues of S
$f_1, \dots, f_p$	Sample eigenvectors of S
D^2	Population Mahalanobis distance (PMD)
T^2	Sample Mahalanobis distance (SMD)
CT^2	Corrected SMD
t_i^2	i th principal component term of T^2
d_i^2	i th principal component term of D^2
ct_i^2	i th principal component term of CT^2
$E[*]$	Expectation of $*$
$V[*]$	Variance of $*$
z_i	Random variable that follows $N(0, 1)$
a_{ij}	Small random coefficient of ϕ_j for f_i
C_i	Coefficient to normalize f_i
l_i^*	Stein's estimator of l_i
ε_F	Relative error of the real variables
ε	Relative error of y
α	Significant level
δ	Non-centrality of generated y
Δ	Non-centrality vector of generated y
n_Σ	Number of generated Σ
n_y	Number of generated y for each Σ
$(\alpha + \beta)_{min}$	Misrecognition rate
Θ	Threshold of $(\alpha + \beta)_{min}$
$p(x \theta)$	Probability density of GMM
w_j	Probability of j th component of GMM
m_j	Average of j th component of GMM
S_j	Covariance matrix of j th component of GMM

B Procedure of simulation

B.1 Monte Carlo simulation of MDs

Monte Carlo simulations were used to generate the MD samples. They were conducted using Microsoft Excel 2010 or

2013 with a program written in Excel Visual Basic for Applications (Excel VBA), and the Excel spreadsheets were only used to demonstrate and visualize the output values. The real variables in Excel VBA have approximately 15 decimal digits of double-precision floating-point real types (IEEE 64-bit). The algorithm of the Monte Carlo simulation is the existing one. However, correcting T^2 by $E[t_i^2]$, which is shown as CT^2 proposed in Sect. 2.2, is new. Algorithm 1 shows the Monte Carlo simulation algorithm, and the details of the input, output, and steps 1) to 14) of Algorithm 1 are as follows:

Algorithm 1 Monte Carlo simulation.

Input: $p, n, n_\Sigma, n_y, \lambda_1, \dots, \lambda_p$, and δ .

Output: PDC, expectation, variance, Hellinger distance from D^2 , misrecognition rate $(\alpha + \beta)_{\min}$, and corresponding threshold Θ of D^2 , T^2 , and CT^2 , respectively.

Do n_Σ times

Generate $\phi_1, \dots, \phi_p$ randomly and define Σ by applying λ_i and ϕ_istep 1)

Generate learning samples $x_1, \dots, x_n$ following $N_p(0, \Sigma)$ to calculate Sstep 2)

Decompose S into l_i and f_i ($i = 1, \dots, p$).step 3)

Generate $\Delta = (\delta_1, \dots, \delta_p)'$ randomly to satisfy $\sum_{i=1}^p \delta_i^2 = \delta$step 4)

Do n_y times

Generate z following $N_p(0, I)$ for D^2 by applying $D^2 = (z + \Delta)^2$step 5)

Generate y following $N_p(\Delta, \Sigma)$ by applying $y = (\phi_i)' \text{diag}(\sqrt{\lambda_i})(z + \Delta)$step 6)

Calculate $t_i = (y - \bar{x}) \cdot f_i / \sqrt{l_i}$ ($i = 1, \dots, p$).step 7)

Calculate T^2 for y by applying $T^2 = \sum_{i=1}^p t_i^2$step 8)

Calculate CT^2 for y by applying $CT^2 = \sum_{i=1}^p t_i^2 / E[t_i^2]$step 9)

Generate the histograms of D^2 , T^2 , and CT^2step 10)

End Do.

End Do.

Calculate the empirical PDCs of D^2 , T^2 , and CT^2 from the histograms.step 11)

Calculate the expectation and variance of the MD. step 12)

Calculate the Hellinger distances of the MDs from D^2step 13)

Calculate the misrecognition rate $(\alpha + \beta)_{\min}$ and the corresponding threshold Θstep 14)

Return.

Input: Predefine $\lambda_1, \dots, \lambda_p$, the population eigenvalues of Σ . All λ_i satisfy $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$. In order to avoid numerical errors in T^2 , all $\lambda_1, \dots, \lambda_p$ must satisfy the condition (43), where ε_F is the relative error of the real variables ($\varepsilon_F \cong 10^{-16}$ for Excel VBA), ε is the relative error of the test sample vector y of T^2 , and $F_\alpha(p, n-p)$ is the upper $100\alpha\%$ point of the F distribution with p and $n-p$ degrees of freedom [40]. (43) is a type of regulation for the condition number of Σ . In this study, it is assumed that $\varepsilon = \varepsilon_F$ and $\alpha = 0.05$. Moreover, the non-centrality δ is predefined.

$$\frac{1}{2} \left(\sqrt{\frac{\lambda_p}{\lambda_1}} - \varepsilon_F \right) \left\{ \frac{n-p-2}{n-p} F_\alpha(p, n-p) - 1 \right\} > \varepsilon_F + \varepsilon. \quad (43)$$

- (1) The population eigenvectors of Σ , $\phi_1, \dots, \phi_p$ (corresponding to $\lambda_1, \dots, \lambda_p$) are generated. All $\phi_1, \dots, \phi_p$ are generated randomly and are orthogonal and normalized. Σ is defined as $\Sigma = \Phi \text{diag}(\lambda_1, \dots, \lambda_p) \Phi'$, where $\Phi = (\phi_1, \dots, \phi_p)$. The total number of generated Σ is n_Σ .
- (2) The learning samples $x_1, \dots, x_n$ following $N_p(0, \Sigma)$ are generated. The sample covariance matrix S is calculated as $S = \sum_{i=1}^n (x_i - \bar{x})^2 / (n-1)$. The total number of generated S is also n_Σ .
- (3) S is decomposed into the sample eigenvalues l_i and sample eigenvectors f_i ($i = 1, \dots, p$). All l_i satisfy $l_1 \geq l_2 \geq \dots \geq l_p$, and all $f_1, \dots, f_p$ are orthogonal and normalized.
- (4) If $\delta > 0$ for generating non-central test samples, $\Delta = (\delta_1, \dots, \delta_p)'$ is generated randomly to satisfy $\sum_{i=1}^p \delta_i^2 = \delta$.
- (5) A population principal component score z is generated, which follows $N_p(0, I)$, i.e., $z = (z_1, \dots, z_p)'$ and $z_i \sim N(0, 1)$ independently. D^2 is calculated as $D^2 = (z + \Delta)^2$. The total number of generated z is $n_\Sigma \cdot n_y$, where n_y is the number of generated y for each generated Σ . It has been confirmed with Q-Q plots prior to these experiments that the simulated D^2 follows a central chi-squared distribution with p degrees of freedom $\chi^2(p)$ if $\delta = 0$ and a non-central chi-squared distribution with p degrees of freedom and non-centrality δ , $\chi^2(p; \delta)$ if $\delta > 0$.
- (6) y that follows $N_p(\Delta, \Sigma)$ with z is generated as $y = \Phi' \text{diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})(z + \Delta)$. The total number of generated y is also $n_\Sigma \cdot n_y$.
- (7) $t_i = (y - \bar{x}) \cdot f_i / \sqrt{l_i}$ ($i = 1, \dots, p$) is calculated, where t_i is the studentized principal component score, i.e., the principal component of T^2 divided by the corresponding sample standard deviation.

- (8) T^2 is calculated for y as $T^2 = \sum_{i=1}^p t_i^2$. The total number of generated T^2 is also $n_\Sigma \cdot n_y$. It has been confirmed with Q-Q plots prior to these experiments that the simulated T^2 follows a central F distribution with p and $n - p$ degrees of freedom, i.e., $F(p, n - p)$, if $\delta = 0$ and a non-central F distribution with the same degrees of freedom, i.e., $F(p, n - p; \delta)$, if $\delta > 0$.
- (9) The corrected T^2 , i.e., CT^2 calculated for y as $CT^2 = \sum_{i=1}^p t_i^2 / E[t_i^2] = \sum_{i=1}^p ct_i^2$, where ct_i is the corrected studentized principal component score, i.e., the principal component of CT^2 is divided by the corresponding sample standard deviation. The total number of each generated CT^2 is also $n_\Sigma \cdot n_y$. All the MDs considered, i.e., D^2 , T^2 , and CT^2 , are listed in Table 1.
- (10) The histograms of all the MDs considered are prepared. In order to compare the MDs of different dimensionality p simultaneously, each sample of the MDs is divided by p for normalization. The number of each class for the sample y of all the normalized MDs is counted. In this study, the width of each class is fixed as 0.01 and the range of all the classes is from 0 to 20. If necessary, the histograms of t_i or ct_i are prepared with the aforementioned width and range.
- (11) The empirical probability density curves (PDCs) of all the MDs are calculated from the histograms.
- (12) The expectation and variance of all the MDs considered are calculated using the histograms.
- (13) The Hellinger distance (HL) of each MD is calculated from D^2 . The HL between two distributions of probability density functions $f(x)$ and $g(x)$ is defined as

$$HL = 2 \int_0^\infty \left(\sqrt{f(x)} - \sqrt{g(x)} \right)^2 dx. \quad (44)$$

- (14) The misrecognition rate $(\alpha + \beta)_{\min}$ and the corresponding threshold Θ between the two MDs corrected by the same method of $\delta = 0$ (central) and $\delta > 0$ (non-central) are calculated. $(\alpha + \beta)_{\min}$ is defined as the minimal sum of α , the type-I error rate of the central distribution, and β , the type-II error rate of the non-central distribution, at the MD value Θ at which the PDCs of the distributions intersect [41].

B.2 EM algorithm

The EM algorithm used in this work is the existing one [36]. However, correcting T^2 by $E[t_i^2]$, as proposed in Sect. 2.2 to improve calculating $g(x|m_j, S_j)$ in (41), is new.

Algorithm 2 EM algorithm for estimating GMM parameters with k components given n independent and identically distributed samples $x_1, x_2, \dots, x_n \in R^p$ [36].

Initialization: $x_1, \dots, x_n$ are divided into k components by the k -means algorithm. The average m_j and covariance matrix S_j are calculated for each component j .

E-step: For $j = 1, \dots, k$, compute at the t th iteration

$$\gamma_{ij}^{(t)} = \frac{w_j^{(t)} g(x_i|m_j^{(t)}, S_j^{(t)})}{\sum_{j=1}^k w_j^{(t)} g(x_i|m_j^{(t)}, S_j^{(t)})}, \quad (i = 1, \dots, n),$$

$$\text{and } n_j^{(t)} = \sum_{i=1}^n \gamma_{ij}^{(t)}.$$

M-step: For $j = 1, \dots, k$, compute the new estimates at the t th iteration

$$w_j^{(t+1)} = \frac{n_j^{(t)}}{n},$$

$$m_j^{(t+1)} = \frac{1}{n_j^{(t)}} \sum_{i=1}^n \gamma_{ij}^{(t)} x_i,$$

$$S_j^{(t+1)} = \frac{1}{n_j^{(t)}} \sum_{i=1}^n \gamma_{ij}^{(t)} (x_i - m_j^{(t+1)})(x_i - m_j^{(t+1)})'.$$

Convergence check: Compute the new log-likelihood $l^{(t+1)}$ and return to E-step if $|l^{(t+1)} - l^{(t)}| > \delta$ for a preset threshold δ ; otherwise end the algorithm.

References

1. Mahalanobis, P.C.: On tests and measures of group divergence. J. Proc. Asiat. Soc. Bengal **26**, 541–588 (1930)
2. Hotelling, H.: The generalization of Student's ratio. Ann. Math. Stat. **2**, 360–378 (1931)
3. Patterson, M.T., Anderson, N., Bennett, C., Bruggemann, J., Grossman, R.L., Handy, M., Ly, V., Mandl, D.J., Pederson, S., Pivarski, J., Powell, R., Spring, J., Wells, W., Xia, J.: The Matsu Wheel: a reanalysis framework for Earth satellite imagery in data commons. Int. J. Data Sci. Anal. **4**, 251–264 (2017)
4. Cudney, E.A., Hong, J., Jugulum, R., Paryani, K., Ragsdell, K.M., Taguchi, G.: An evaluation of Mahalanobis-Taguchi system and neural network for multivariate pattern recognition. J. Ind. Syst. Eng. **1**, 139–150 (2007)
5. Su, C.-T., Wang, P.-C., Chen, Y.-C., Chen, L.-F.: Data mining techniques for assisting the diagnosis of pressure ulcer development in surgical patients. J. Med. Syst. **36**, 2387–2399 (2012)
6. Ghasemi, E., Aghaie, A., Cudney, E.A.: Mahalanobis Taguchi system: a review. Int. J. Qual. Reliab. Manag. **32**, 291–307 (2015)
7. Mota-Gutiérrez, C.G., Reséndiz-Flores, E.O., Reyes-Carlos, Y.I.: Mahalanobis-Taguchi system: state of the art. Int. J. Qual. Reliab. Manag. **35**, 596–613 (2018)
8. Mukhopadhyay, P.: Multivariate Statistical Analysis. World Scientific, Singapore (2009)

9. Tracy, N.D., Young, J.C., Mason, R.L.: Multivariate control charts for industrial observations. *J. Qual. Technol.* **24**, 88–95 (1992)
10. Kimura, F., Takashina, K., Tsuruoka, S., Miyake, Y.: Modified quadratic discriminant functions and the application to Chinese character recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **PAMI-9**(1), 149–153 (1987)
11. Friedman, J.H.: Regularized discriminant analysis. *J. Am. Stat. Assoc.* **84**, 165–175 (1989)
12. Sun, F., Omachi, S., Aso, H.: Precise selection of candidates for handwritten character recognition using feature regions. *IEICE Trans. Inf. Syst.* **E79-D**, 510–515 (1996)
13. Fauvel, M., Chanussot, J., Benediktsson, J.A., Villa, A.: Parsimonious Mahalanobis kernel for the classification of high dimensional data. *Pattern Recognit.* **46**, 845–854 (2013)
14. Sakai, M., Yoneda, M., Hase, H., Maruyama, H., Naoe, M.: A quadratic discriminant function based on bias rectification of eigenvalues. *IEICE Trans. Inf. Syst.* **J82-D-II**, 631–640 (1999). (in Japanese)
15. Iwamura, M., Omachi, S., Aso, H.: Estimation of true Mahalanobis distance from eigenvectors of sample covariance matrix. *IEICE Trans. Inf. Syst.* **J86-D-II**, 22–31 (2003). (in Japanese)
16. Jorgensen, T., Rothrock, R.: Correcting for bias in Mahalanobis and log-likelihood estimates. *IEEE Trans. Aerosp. Electron. Syst.* **46**, 2078–2089 (2010)
17. Oehlert, G.W.: A note on the delta method. *Am. Stat.* **46**(1), 27–29 (1992)
18. Kobayashi, Y.: A proposal of simple correcting scheme for sample Mahalanobis distances using delta method. *IEICE Trans. Inf. Syst.* **J97-D**(8), 1228–1236 (2014). (in Japanese)
19. Kobayashi, Y.: A corrector for the sample Mahalanobis distance free from estimating the population eigenvalues of covariance matrix. In: Hirose, A., et al. (eds.) *ICONIP 2016, Part II. LNCS*, vol. 9948, pp. 224–232. Springer (2016)
20. Kulis, B.: Metric learning: a survey. *Found. Trends Mach. Learn.* **5**, 287–364 (2012)
21. Schultz, M., Joachims, T.: Learning a distance metric from relative comparisons. In: *Advances in Neural Information Processing Systems (NIPS)* (2003)
22. Kwok, J., Tsang, I.: Learning with idealized kernels. In: *Proceedings of International Conference on Machine Learning (ICML)* (2003)
23. Shalev-Shwartz, S., Singer, Y., Ng, A.Y.: Online learning of pseudometrics. In: *Proceedings of International Conference on Machine Learning (ICML)* (2004)
24. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning, with application to clustering with side-information. In: *Advances in Neural Information Processing Systems (NIPS)* (2002)
25. Weinberger, K.Q., Blitzer, J., Saul, L.K.: Distance metric learning for large margin nearest neighbor classification. In: *Advances in Neural Information Processing Systems (NIPS)* (2005)
26. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: *Advances in Neural Information Processing Systems (NIPS)* (2004)
27. Globerson, A., Roweis, S.: Metric learning by collapsing classes. In: *Advances in Neural Information Processing Systems (NIPS)* (2005)
28. Davis, J., Kulis, B., Jain, P., Sra, S., Dhillon, I.: Information-theoretic metric learning. In: *Proceedings of International Conference on Machine Learning (ICML)* (2007)
29. Lawley, D.N.: Tests of significance for the latent roots of covariance and correlation matrices. *Biometrika* **43**, 128–136 (1956)
30. Girshick, M.A.: On the sampling theory of the roots of determinantal equations. *Ann. Math. Stat.* **10**, 203–224 (1939)
31. Muirhead, R.J.: *Aspects of Multivariate Statistical Theory*. Wiley, New York (1982)
32. Anderson, T.W.: *An Introduction to Multivariate Statistical Analysis*, 3rd edn. Wiley, New York (2003)
33. Takemura, A., Sheena, Y.: Distribution of eigenvalues and eigenvectors of Wishart matrix when the population eigenvalues are infinitely dispersed and its application to minimax estimation of covariance matrix. *J. Multivar. Anal.* **94**(2), 271–299 (2005)
34. Takemura, A.: A principal decomposition of Hotelling's T^2 statistic. In: Krishnaiah, P.R. (ed.) *Multivariate Analysis-VI*, pp. 583–597. Elsevier, Amsterdam (1985)
35. James, W., Stein, C.: Estimation with quadratic loss. In: *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379 (1961)
36. Gupta, M.R., Chen, Y.: Theory and use of the EM algorithm. *Found. Trends Signal Process.* **4**(3), 223–296 (2010)
37. Dua, D., Karra Taniskidou, E.: UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA. <http://archive.ics.uci.edu/ml> (2017). Accessed 27 Apr 2018
38. McLachlan, G.J., Peel, D.: *Finite Mixture Models*. Wiley, New York (2000)
39. Vidal, R., Ma, Y., Sastry, S.S.: *Generalized Principal Component Analysis*. Springer, New York (2016)
40. Kobayashi, Y.: Effects of numerical errors on sample Mahalanobis distances. *IEICE Trans. Inf. Syst.* **E99-D**(5), 1337–1344 (2016)
41. Fukunaga, K.: *Introduction to Statistical Pattern Recognition*, 2nd edn. Academic Press, New York (1990)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.