PAPER

Effects of Numerical Errors on Sample Mahalanobis Distances

Yasuyuki KOBAYASHI[†a)], ***Member***

SUMMARY The numerical error of a sample Mahalanobis distance ($T^2 = y'\mathrm{S}^{-1}y$) with sample covariance matrix S is investigated. It is found that in order to suppress the numerical error of T^2, the following conditions need to be satisfied. First, the reciprocal square root of the condition number of S should be larger than the relative error of calculating floating-point real-number variables. The second proposed condition is based on the relative error of the observed sample vector y in T^2. If the relative error of y is larger than the relative error of the real-number variables, the former governs the numerical error of T^2. Numerical experiments are conducted to show that the numerical error of T^2 can be suppressed if the two above-mentioned conditions are satisfied.

key words: *sample Mahalanobis distance, numerical error, condition number, round-off error*

1. Introduction

When the Mahalanobis distance is applied to statistical machine learning, Hotelling's T^2 statistic is commonly used to measure the sample Mahalanobis distance (sample MD) between two populations. Let $x_1, \cdots, x_n$ and y be n learning sample vectors and an observed sample vector independently following a p-variate normal distribution $\mathrm{N}_p(\mu, \Sigma)$ with mean μ and covariance matrix Σ, and let the mean vector $\bar{x}$ and the sample covariance matrix S be calculated from $x_1, \cdots, x_n$. Then, the sample MD for the observed sample vector y is defined by $T^2 = (y - \bar{x})'\,\mathrm{S}^{-1}\,(y - \bar{x})$.

The sample MD cannot avoid statistical fluctuation in x_i and y, given by variance $V[T^2]$. In addition, it cannot avoid the effect of the computer-based numerical error ΔT^2. To the best of my knowledge, no study has discussed the effects of numerical errors on T^2 thus far, whereas their effects on the eigenvalues and eigenvectors of S have been discussed previously (see [1] and [2]). For example, rounding errors can make the zero roots of Σ positive [3]. Rounding errors have a small influence on the eigenvectors of S [4], [5], and this effect can be bounded by the eigenvalues of Σ [4] (see [1] and [2]). The condition number cond (S), i.e., the ratio of the maximum eigenvalue of S to the minimum eigenvalue of S, is said to be the index of ΔT^2 [6]. However, cond (S) is actually the index of only S, and not of the sample MD, because the condition number is originally derived from the formula for estimating the relative error of simultaneous linear equations. When a system of simultaneous linear equations $\mathrm{A}x = f$ has the solution x, the upper limit of an error Δx in x is given by (1),

$$\frac{||\Delta x||}{||x||} \leq \frac{\mathrm{cond}\,(\mathrm{A})}{1 - \dfrac{||\Delta \mathrm{A}||}{||\mathrm{A}||}\mathrm{cond}\,(\mathrm{A})} \left\{ \frac{||\Delta f||}{||f||} + \frac{||\Delta \mathrm{A}||}{||\mathrm{A}||} \right\}, \tag{1}$$

which includes cond (A), where it is assumed that a vector norm $||f||$ is the quadratic norm given by $||f|| = \sqrt{\sum_{i=1}^{p} f_i^2}$ and that the matrix norm $||\mathrm{A}||$ is the spectral norm given by the maximum eigenvalue of A if A is a real square matrix, or $\sqrt{\mathrm{A}'\mathrm{A}}$ if A is not [7]. These norm definitions are applied throughout this paper.

Furthermore, one of the reasons for the difficulty in understanding the estimation of the fluctuation in the sample MD is the coexistence of the statistical fluctuation (i.e., $V[T^2]$) and the numerical fluctuation (i.e., ΔT^2). Only the statistical fluctuation can be described by the F-distribution that the sample MD follows. A previous study [2] proposed a probability distribution that the absolute round-off error of floating-point real-number type follows [8]. However, the probability distribution of the round-off error shares a complicated relationship with the distribution of the true value [8]; therefore, it is difficult to describe the probability distribution of the sample MD with both the above-mentioned fluctuations. Therefore, sufficient computer precision is required to suppress ΔT^2. For example, it has been reported that the effect of ΔT^2 is sufficiently small with a computer precision of around 15 decimal digits of double-precision real type when the ratio of the maximum eigenvalue to the minimum eigenvalue of the covariance matrix Σ is at most 4 decimal digits [6]. However, in Ref. [6], the above reason was not shown theoretically. As shown in this paper theoretically, ΔT^2 can be suppressed even if cond (S) is much larger than that in Ref. [6]. It is necessary to show the conditions for suppressing ΔT^2 by theoretical investigation of the sample MD.

When a classifier is applied to statistical machine learning, the classification performance of the classifier is the most important factor. Therefore, it is crucial to obtain the conditions for suppressing the effect of ΔT^2 for classification performance in statistical machine learning.

In short, this paper shows that the numerical error ΔT^2 of the sample MD is governed by cond (S) and the relative errors of both the calculated floating-point real-number variables and the observed sample vector y; the latter is not men-

Manuscript received August 31, 2015.
Manuscript revised December 26, 2015.
Manuscript publicized February 12, 2016.

[†]The author is with the Faculty of Science and Engineering, Teikyo University, Utsunomiya-shi, 320–8551 Japan.
a) E-mail: ykoba@ics.teikyo-u.ac.jp

DOI: 10.1587/transinf.2015EDP7348

tioned in Ref. [6]. Further, this paper presents the conditions for suppressing the effect of ΔT^2. The availability of these conditions is investigated through numerical experiments.

The remainder of this paper is organized as follows. Section 2 describes formulae for estimating the effect of both cond (S) and the computer precision (e.g., relative error of floating-point arithmetic). In addition, methods for evaluating the numerical errors are discussed and the condition for suppressing the effect of ΔT^2 in the sample MD is proposed. Section 3 presents the results of numerical experiments conducted to investigate the relationship between ΔT^2 and the classification performance of the sample MD. Finally, Sect. 4 summarizes my findings and concludes the paper.

2. Theoretical Investigation of Numerical Error and Classification Performance of a Sample MD

2.1 Formula for Numerical Error of a Sample MD

First, the upper limit of the relative numerical error $\left|\Delta T^2\right| / T^2$ will be shown. Let $x_1, \cdots, x_n$ and y be n learning sample vectors and an observed sample vector independently following a p-variate normal distribution $\mathrm{N}_p(\mu, \Sigma)$ with mean μ and covariance matrix Σ. Let the mean vector $\bar{x}$ be calculated from $x_1, \cdots, x_n$ and define the data matrix $\mathrm{X} \equiv (x_1 - \bar{x}, \cdots, x_n - \bar{x})'$. Let the sample covariance matrix S be $\mathrm{S} = X'X/(n-1)$ with maximum eigenvalue l_{max} and minimum eigenvalue l_{min}. The upper limit of $\left|\Delta T^2\right| / T^2$ is given by (2) (see the Appendix for details). Parameters η_1 and η_2 in (2) are given by $\eta_1 = \sqrt{l_{max}/l_{min}} \cdot \varepsilon_F$ and $\eta_2 = \sqrt{l_{max}/l_{min}} \cdot \varepsilon$, where ε_F is the relative error of the real variables of T^2 calculation, and $\varepsilon = \|\Delta y\| / \|y - \bar{x}\|$ denotes the relative numerical error of y, the observed sample vector.

$$\frac{\left|\Delta T^2\right|}{T^2} \leq \frac{2}{1 - \sqrt{\frac{l_{max}}{l_{min}}} \cdot \varepsilon_F} \left\{ \sqrt{\frac{l_{max}}{l_{min}}} \cdot \varepsilon_F + \sqrt{\frac{l_{max}}{l_{min}}} \cdot \varepsilon \right\} = \frac{2(\eta_1 + \eta_2)}{1 - \eta_1}. \tag{2}$$

In order to suppress the numerical error ΔT^2, i.e., $\left|\Delta T^2\right| / T^2 \ll 1$, the right-hand side of (2) should be much smaller than 1; hence, $\eta_1 \ll 1$ and $\eta_2 \ll 1$. Here, $\eta_1 \ll 1$ is given by (3) and $\eta_2 \ll 1$ is given by (4):

$$\sqrt{\frac{l_{min}}{l_{max}}} \gg \varepsilon_F, \tag{3}$$

$$\sqrt{\frac{l_{min}}{l_{max}}} \gg \varepsilon. \tag{4}$$

The next objective is to obtain conditions with a confidence probability for suppressing ΔT^2. First, the existence of the norm of $\left|\Delta T^2\right|$ should satisfy $\eta_1 < 1$, and (3) becomes

$$\sqrt{\frac{l_{min}}{l_{max}}} > \varepsilon_F. \tag{5}$$

Second, if the upper limit of $\Delta T^2 / T^2$ in (2) is less than R_{max}, i.e., the upper limit of the ratio of the statistical fluctuation of T^2, ΔT^2 is sufficiently small with a probability of $1 - \alpha$ in comparison with the statistical fluctuation of T^2:

$$\frac{\left|\Delta T^2\right|}{T^2} \leq \frac{2(\eta_1 + \eta_2)}{1 - \eta_1} < R_{max}, \tag{6}$$

where R_{max} is given by

$$R_{max} = \frac{T_\alpha^2 - \mathrm{E}[T^2]}{\mathrm{E}[T^2]}, \tag{7}$$

where T_α^2 is the upper $100\alpha\%$ point of T^2 and $\mathrm{E}[T^2]$ is the expectation of T^2. Here, the sample MD without the numerical error ΔT^2 follows an F-distribution with p and $n-p$ degrees of freedom [1]. This F-distribution is called a central F-distribution, in contrast to a non-central F-distribution, as mentioned later.

$$T^2 = (y - \bar{x})' \, \mathrm{S}^{-1} (y - \bar{x}) = \frac{(n-1)\,p}{n-p} F(p, n-p). \tag{8}$$

From (8), the expectation $\mathrm{E}[T^2]$ is obtained as $\mathrm{E}[T^2] = (n-1)\,p/(n-p-2)$, and (7) is expressed as $R_{max} = (n-p-2)/(n-p) \cdot F_\alpha(p, n-p) - 1$, where F_α is the upper $100\alpha\%$ point of the F-distribution. Therefore, (6) is expressed as

$$\frac{2(\eta_1 + \eta_2)}{1 - \eta_1} < R_{max} = \frac{n-p-2}{n-p} F_\alpha(p, n-p) - 1. \tag{9}$$

If $\eta_1 \ll 1$, the upper limit of (2) can be approximated by $2(\eta_1 + \eta_2)/(1 - \eta_1) \cong 2\eta_2$, and (6) with a probability of $1 - \alpha$ is simplified as

$$2\eta_2 = 2\varepsilon \sqrt{\frac{l_{max}}{l_{min}}} < R_{max} = \frac{n-p-2}{n-p} F_\alpha(p, n-p) - 1. \tag{10}$$

Therefore, satisfying (5) and (9) is the realistic condition for suppressing ΔT^2. For example, under the condition of the numerical experiment in this paper, where $p = 7$ and $n = 15$, $R_{max} \cong 1.63$ when $\alpha = 5\%$ and (10) becomes $\eta_2 < 0.91$. Thus, $\eta_1 < 1$ and the upper limit of $\Delta T^2 / \left|T^2\right|$: $2(\eta_1 + \eta_2)/(1 - \eta_1) < R_{max} \cong 1.63$, or $\eta_1 \ll 1$ and $\eta_2 < 0.91$ need to be satisfied for ΔT^2 to be sufficiently smaller than the statistical fluctuation with a probability of $1 - \alpha = 95\%$.

2.2 Relationship between Classification Performance and Variance of Probability Distribution

To investigate the classification performance of the sample MD affected by numerical error, the relationship between the classification performance of the sample MD and the variance from the probability density function is considered as follows.

The classification performance can be estimated by

both the distribution of the learning samples (normal distribution) and a distribution that is definitely different from that of the learning samples (abnormal distribution). Let n learning samples follow a p-variate normal distribution with mean μ and covariance matrix Σ, $\mathrm{N}_p(\mu, \Sigma)$. Without any numerical error, the sample MD of the learning samples follows (8), i.e., a central F-distribution. However, it is difficult to depict an abnormal distribution because the population covariance matrix Σ' of an abnormal distribution is generally different from Σ of a normal distribution. If $\Sigma' = \Sigma$, the sample MD of an abnormal distribution follows a non-central F-distribution and is given by

$$\begin{aligned} T'^2 &= (u - \bar{x})' \, \mathrm{S}'^{-1} (u - \bar{x}) \\ &= \frac{(n-1)\,p}{n-p} F(p, n-p; \delta), \end{aligned} \tag{11}$$

where a test sample of the abnormal distribution follows $u \sim \mathrm{N}_p(\mu + m, \Sigma)$, $\bar{x}$ is the sample mean of the learning samples, and the non-centrality $\delta = m'\Sigma^{-1}m$. However, a non-central F-distribution cannot be described by simple expressions because its probability density function is given by a hypergeometric function that is an infinite series. In this study, numerical calculation of a non-central F-distribution was executed using Mathematica.

The case of $\Sigma' = \Sigma$ corresponds to linear discriminant analysis (LDA). LDA can improve the classification performance by estimating the pooled covariance matrix among the classes [9]. This paper considers only $\Sigma' = \Sigma$ (not $\Sigma' \neq \Sigma$) to analyze the classification performance of the sample MD as theoretically as possible.

In Fig. 1, the central F-distribution represents the normal distribution to which the learning samples belong, and the non-central F-distribution represents the abnormal distribution. For a selected threshold Θ below which a sample belongs to the normal distribution, the probability of the normal distribution $\alpha = \mathrm{P}_c(X \geq \Theta)$ is called the type I error probability, and the probability of the abnormal distribution $\beta = \mathrm{P}_n(X \leq \Theta)$ is called the type II error probability. The classification performance of the sample MD corresponds to $(\alpha + \beta)_{min}$ with the threshold Θ minimizing $\alpha + \beta$ (misclassification probability). In Fig. 1, the threshold Θ is the MD

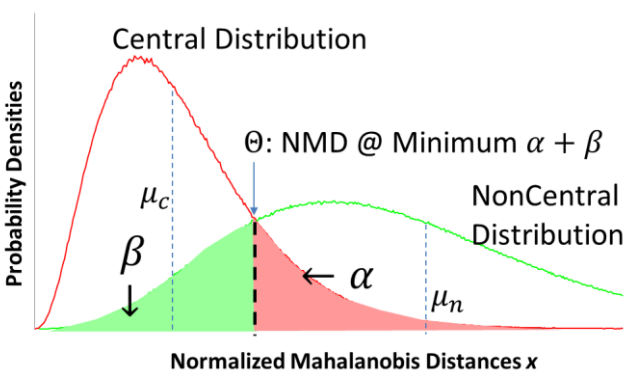


Fig. 1 Relationship between minimum misclassification probability $\alpha+\beta$ and probability distribution.

value located at the intersection point of two single-peak distributions [10]. Therefore, using both the normal distribution and the abnormal distribution with non-centrality δ, the classification performance $(\alpha + \beta)_{min}$ and the threshold Θ are obtained.

Here, the relationship among the upper limit of $(\alpha + \beta)_{min}$ and the variances of the two distributions will be investigated. For the central distribution with expectation $\mu_c < \Theta$ and variance V_c in Fig. 1, the Markov inequality $\mathrm{P}_c(|X| \geq \Theta) \leq \mathrm{E}[X^2]/\Theta^2$ is obtained and $\alpha = \mathrm{P}_c(X \geq \Theta) \leq \left(\mu_c^2 + V_c\right)/\Theta^2$ because the central distribution is defined by $x > 0$. For the non-central distribution with expectation $\mu_n > \Theta$ and variance V_n in Fig. 1, the Chebyshev inequality $\mathrm{P}_n(|X - \mu_n| \geq |\mu_n - \Theta|) \leq V_n/(\mu_n - \Theta)^2$ is obtained and $\beta = \mathrm{P}_n(X \leq \Theta) \leq V_n/\left\{2(\mu_n - \Theta)^2\right\}$ because the right tail area of the non-central distribution is larger than the left tail area. Therefore, the upper limit of $(\alpha + \beta)_{min}$ in Fig. 1 is given by

$$(\alpha + \beta)_{min} \leq \frac{\mu_c^2 + V_c}{\Theta^2} + \frac{V_n}{2(\mu_n - \Theta)^2}. \tag{12}$$

If V_c and V_n increase because of numerical error, $(\alpha + \beta)_{min}$ may increase and the classification performance may deteriorate.

3. Numerical Experiments

3.1 Procedure

The numerical experiments in this study were conducted using Microsoft Excel 2010 with a source code written in Excel VBA. The algorithm of the source code follows steps 1–9 as described below. Real variables in Excel VBA have around 15 decimal digits of double-precision real type (IEEE 64-bit); hence, the relative error ε_F is given by $\varepsilon_F \cong 10^{-16}$. Table 1 lists the common parameters used in all the experiments.

1) $\lambda_1 > \cdots > \lambda_p$, i.e., the eigenvalues of the p-variate population covariance matrix Σ, are initially defined. Each $\sqrt{\lambda_i}$ has a value given by a geometric sequence between the first term $\sqrt{\lambda_{max}} = \sqrt{\lambda_1}$ and the last term $\sqrt{\lambda_{min}} = \sqrt{\lambda_p}$.
2) To estimate the dependence of $(\alpha + \beta)_{min}$ on the non-centrality parameter, δ is increased from 5 to 30 in increments of 5, and steps 3–9 are executed at each δ value.
3) The population eigenvectors $\phi_1, \cdots, \phi_p$ of Σ are generated by random numbers. A non-centrality vector

Table 1 Common parameters.

Dimensionality	$p = 7$
Number of learning samples	$n = 15$
Number of test samples	$m = 1000$
Number of times covariance matrices are generated	$k = 10000$
Non-centrality	$\delta = 5$~30 (increment 5)

$\Delta = \left(\delta_1, \cdots, \delta_p\right)$ is generated by random numbers to satisfy $\delta = \delta_1^2 + \cdots + \delta_p^2$.

4) Central learning samples x_i and non-central learning samples $\tilde{x}_i$ $(i = 1, \cdots, n)$ are generated by p-variate normal random numbers, x_i, $\tilde{x}_i = \mathrm{N}_p(0, \Sigma)$. The sample eigenvalues and eigenvectors l_i, v_i $(i = 1, \cdots, p)$ are obtained by the sample covariance matrix S of x_i. The other sample eigenvalues and eigenvectors $\tilde{l}_i$, $\tilde{v}_i$ $(i = 1, \cdots, p)$ are obtained by the sample covariance matrix of $\tilde{x}_i$.
5) The maximum numerical error ε_A of the eigenvalues is calculated as $\varepsilon_A = \max_{i=1,\cdots,p} \left\| \mathrm{S}^{1/2} v_i - \sqrt{l_i} v_i \right\|$. After the numerical experiments, the relative error of the real variables $\varepsilon_F \cong 10^{-16}$ and the value of $\sqrt{\lambda_{max}}$ satisfies $\varepsilon_A \cong \sqrt{l_{max}} \cdot \varepsilon_F$ (see the Appendix for details).
6) Central test samples y_i and non-central test samples $\tilde{y}_i$ $(i = 1, \cdots, m)$ are generated by $y_i = \mathrm{N}_p(0, \Sigma)$ and $\tilde{y}_i = \mathrm{N}_p(\Delta, \Sigma)$.
7) In order to control the relative error of y artificially, each element of the test sample vectors y_i and $\tilde{y}_i$ is operated by editing its mantissa as the relative error limit of the element ε_e is selected from among 10^{-6}, 10^{-10}, and, 10^{-16} $(\cong \varepsilon_F)$, i.e., the default value of the real variable. If $\varepsilon_e > \varepsilon_F$, lower values of the mantissa corresponding to values lower than ε_e are changed to 0 by rounding them off. The real variable of each element of y_i and $\tilde{y}_i$ having relative error ε_F becomes the effective relative error of the element ε_e. According to Ref. [4], let the round-off errors have the common assumption that they are uniformly distributed, independent of each other, and independent of the unrounded value. Then, the expectation of the relative error of $\mathrm{E}\left[\|\Delta y\| / \|y\|\right] = \sqrt{p/3} \cdot \varepsilon_e$. Under the above range of ε_e, $\|\Delta y\| / \|y\| \cong \|\Delta y\| / \|y - \bar{x}\|$.
8) For the central test sample y_i and the non-central test sample $\tilde{y}_i$, Mahalanobis distances are calculated and normalized by the dimensionality as $CSMD_i = \sum_{j=1}^{p} \{(y_i - \bar{x}) \cdot v_i\}^2 / \left(l_j \cdot p\right)$ and $NSMD_i = \sum_{j=1}^{p} \{(\tilde{y}_i - \bar{\tilde{x}}) \cdot \tilde{v}_i\}^2 / \left(\tilde{l}_j \cdot p\right)$.
9) Steps 2–8 are repeated k times and a total of $k \cdot m =$ 10,000,000 $CSMD_i$ and $NSMD_i$ samples are obtained. By using these samples, the following estimates are calculated.

- The expectation and variance of $CSMD_i$ and $NSMD_i$.
- From the histograms of $CSMD_i$ and $NSMD_i$ samples, Q-Q plots are obtained. A Q-Q plot is drawn as follows. (i) The empirical cumulative distribution function $F(x)$ is generated by the MD histogram from all the MD samples. (ii) The cumulative distribution function $G(y)$ is calculated by the theoretical expression for comparison with $(n-1)/(n-p) \cdot F(p, n-p)$ for $CSMD_i$ or $(n-1)/(n-p) \cdot F(p, n-p; \delta)$ for $NSMD_i$. (iii) All the MD samples obtained by the numerical experiment are plotted. The x-axis value of each point is the MD value x obtained by all the MD samples and the corresponding y-axis value is given by $y = F^{-1}(G(x))$. (iv) If all the points are on the line $y = x$, the MD samples follow the theoretical probability distribution used for comparison.
- In order to quantify the discrepancy of the points on the Q-Q plot from the line $y = x$, the slope of the regression line intersecting the origin for the points on the Q-Q plot with $CSMD_i < 5$ or $NSMD_i < 5$ is calculated. The ranges correspond with 99.999% of the left tail probability of all the MD samples.
- The minimum misclassification probability $(\alpha + \beta)_{min}$ is calculated using the histograms of both $CSMD_i$ and of $NSMD_i$.

3.2 Results

This subsection discusses the results of the numerical experiments described in the previous subsection.

First, the phenomenon whereby the numerical error of the sample Mahalanobis distance T^2 affects the probability distribution and increases the variance of T^2 was investigated. Table 2 summarizes the calculated numerical error parameters concerning the upper limit of $\left|\Delta T^2\right| / T^2$ in (2) under three typical cases. Note that η_1 and η_2 are calculated using the population eigenvalues λ because the sample eigenvalues l fluctuate in each trial of the numerical experiment and the expectation $\mathrm{E}[l]$ is affected by numerical errors. For example, the minimum sample eigenvalue l_{min} is lower than the minimum population eigenvalue λ_{min} [11] without any numerical error; however, owing to numerical errors, $\mathrm{E}[l_{min}]$ is larger than λ_{min} in Case B, in contrast to

Table 2 Typical cases and relative numerical error parameters.

Case	A	B	C
$\sqrt{\lambda_{max}}$	10^0		
$\sqrt{\lambda_{min}}$	10^{-12}	10^{-18}	10^{-12}
ε	$10^{-16} (\cong \varepsilon_F)$		10^{-10}
$\eta_1 = \frac{\varepsilon_A}{\sqrt{\lambda_{min}}}$	3×10^{-4}	3×10^{2}	3×10^{-4}
$\eta_2 = \sqrt{\frac{\lambda_{max}}{\lambda_{min}}}\varepsilon$	1×10^{-4}	1×10^{2}	1×10^{2}
$\frac{2(\eta_1+\eta_2)}{1-\eta_1}$	8×10^{-4}	Impossible	2×10^{2}
$\eta_1 < 1$	True	False	True
$\frac{2(\eta_1+\eta_2)}{1-\eta_1} < R_{max}$	True	False	False
$\sqrt{\mathrm{E}[l_{max}]}$	1×10^{0}	1×10^{0}	1×10^{0}
$\sqrt{\mathrm{E}[l_{min}]}$	8×10^{-13}	5×10^{-17}	8×10^{-13}
ε_A	3×10^{-16}	3×10^{-16}	3×10^{-16}
$\tilde{\eta}_1 = \frac{\varepsilon_A}{\sqrt{\mathrm{E}[l_{min}]}}$	4×10^{-4}	6×10^{0}	4×10^{-4}
$\tilde{\eta}_2 = \sqrt{\frac{\mathrm{E}[l_{max}]}{\mathrm{E}[l_{min}]}}\varepsilon$	1×10^{-4}	2×10^{0}	1×10^{2}

Ref. [11]. This is because the matrix calculation is affected by the numerical error under $\sqrt{\lambda_{min}} < \varepsilon_F \cong 10^{-16}$. Therefore, η_1 and η_2 are calculated using $\sqrt{\lambda}$ (not $\sqrt{\mathrm{E}[l]}$).

As discussed in Sect. 2.1, the suppression condition that $\eta_1 < 1$ and the upper limit of $\Delta T^2/\left|T^2\right|$: $2(\eta_1+\eta_2)/(1-\eta_1) < R_{max} \cong 1.63$, or $\eta_1 \ll 1$ and $\eta_2 < 0.91$ need to be satisfied for the numerical error ΔT^2 to be sufficiently smaller than the statistical fluctuation with the probability of $1-\alpha = 95\%$. In Case A, which satisfies (5) and (9), the numerical error will be negligible because $\eta_1 \ll 1$ and $2(\eta_1+\eta_2)/(1-\eta_1) \ll 1$. However, in Case B, which does not satisfy (5) and (9), the numerical error will not be negligible because $\eta_1 \gg 1$. Further, in Case C, which satisfies (5) but not (9), the numerical error will not be negligible because $\eta_1 \ll 1$ and $2(\eta_1+\eta_2)/(1-\eta_1) \gg 1$. Therefore, the probability distribution of the sample MD in only Case A can be shown by the central F-distribution or non-central F-distribution theoretically.

In Fig. 2, each tail of the central and non-central distributions in Cases B and C is broader than that in Case A; hence, the variances of Cases B and C are larger than that of Case A. Furthermore, in Fig. 3, each Q-Q plot of the central and non-central distributions in Case A nearly coincides with the line $y = x$; hence, both the distributions follow the theoretical distributions. However, both the Q-Q plots of central and non-central distributions in Cases B and C are distinct from the line $y = x$; hence, the distributions in Cases B and C do not follow the theoretical distributions. Therefore, as estimated in Table 2, the probability distributions of the sample MD in Cases B and C are affected by the numerical error.

Second, the minimum misclassification probability $(\alpha+\beta)_{min}$ of the sample MD was investigated. As the non-centrality δ is increases, $(\alpha+\beta)_{min}$ decreases monotonically. The dependence of $(\alpha+\beta)_{min}$ on numerical errors needs to be estimated by the curves of $(\alpha+\beta)_{min}$. In Fig. 4(a), the curves of $(\alpha+\beta)_{min}$ in Cases B and C are beyond that in Case A. Therefore, $(\alpha+\beta)_{min}$ of the sample MD deteriorates if the numerical error is not negligible. In Fig. 4(b), the threshold of the sample MD corresponding to $(\alpha+\beta)_{min}$ in Case C is larger than those in Cases A and B.

Finally, the dependence on various $\sqrt{\lambda_{max}}$, $\sqrt{\lambda_{min}}$, and ε, where the common condition is $p = 7$ and $n = 15$, is summarized in Fig. 5. Under each horizontal axis in Fig. 5, all the cases of $\sqrt{\lambda_{max}}$, $\sqrt{\lambda_{min}}$, and ε are enumerated and tabulated. In Fig. 5(a), η_1, η_2, and the upper limit of the relative error $\left|\Delta T^2\right|/T^2$: $2(\eta_1+\eta_2)/(1-\eta_1)$ are shown. To observe the difference between the cases $\eta_1 \cong \eta_2$ and $\eta_1 < \eta_2$, the values of ε are modified as follows. As $\eta_1 = \sqrt{l_{max}/l_{min}} \cdot \varepsilon_F$ and $\eta_2 = \sqrt{l_{max}/l_{min}} \cdot \varepsilon$, if $\varepsilon = 10^{-16}$ ($\cong \varepsilon_F$), then $\eta_1 \cong \eta_2$; if $\varepsilon = 10^{-10}$ or 10^{-6} ($\gg \varepsilon_F$), then $\eta_1 < \eta_2$. As in the case of the previous results, the numerical error is supposed to be negligible under the suppression condition that $\eta_1 < 1$ and the upper limit of $\Delta T^2/T^2$: $2(\eta_1+\eta_2)/(1-\eta_1) < R_{max} \cong 1.63$, or $\eta_2 < 0.91$ and $\eta_1 \ll 1$ when $p = 7$ and $n = 15$ for

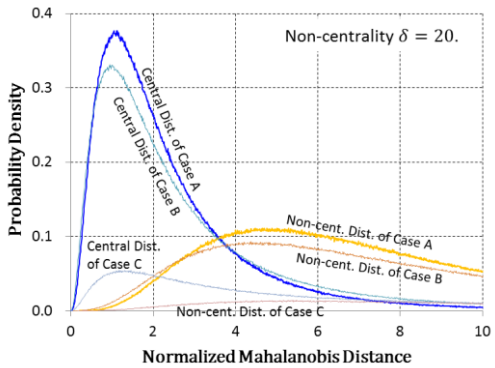


Fig. 2 Dependence of probability density distributions of sample MDs on numerical errors.

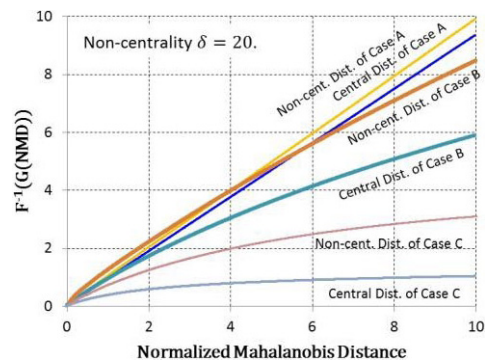
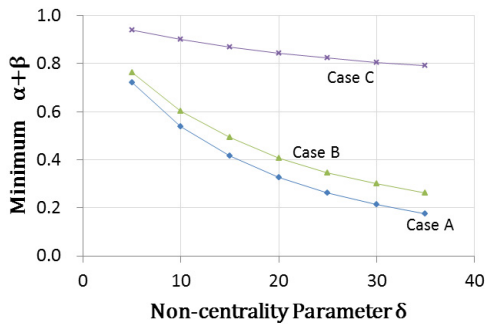
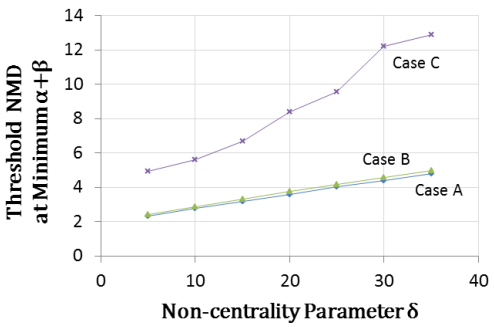


Fig. 3 Dependence of Q-Q plots of sample MDs on numerical errors.

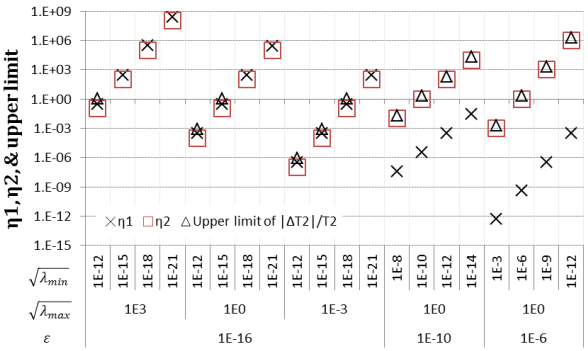


(a) $(\alpha+\beta)_{min}$.

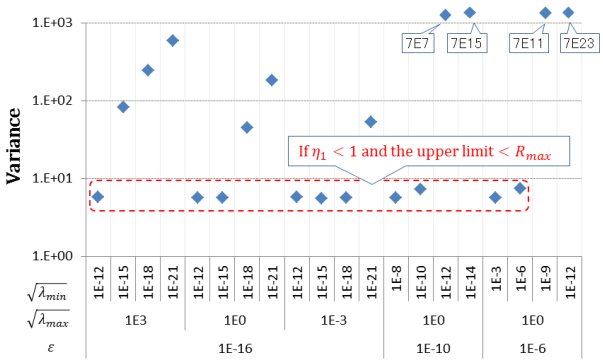


(b) Threshold of $(\alpha+\beta)_{min}$.

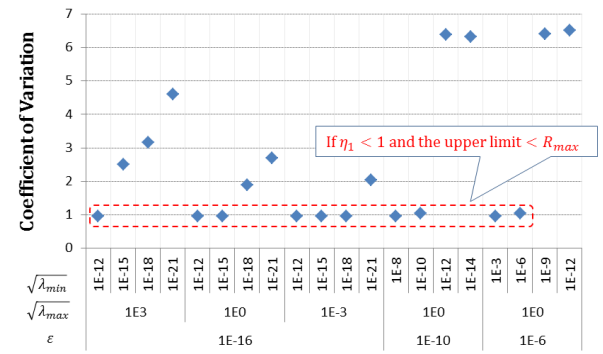
Fig. 4 Dependence of minimum misclassification probability $(\alpha+\beta)_{min}$ of sample MDs on numerical errors.



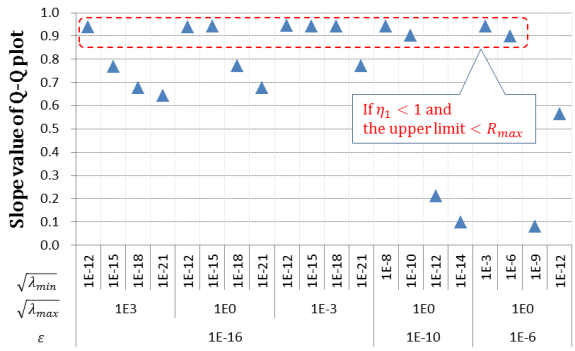
(a) η_1, η_2, and the upper limit of $|\Delta T^2|/T^2$.



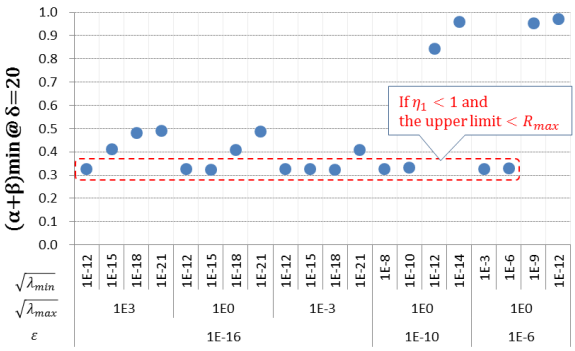
(b) Variances of central distributions.



(c) Coefficients of variation of central distributions.



(d) Slope values of Q-Q plots.



(e) Minimum misclassification probability at non-centrality $\delta = 20$.

Fig. 5 Effects of numerical errors on sample MDs.

ΔT^2 to be sufficiently smaller than the statistical fluctuation with the probability of $1 - \alpha = 95\%$. If $\eta_1 < 1$, the upper limit of $|\Delta T^2|/T^2$ exists and can be calculated. In Fig. 5(b), the variances of normalized sample MDs T^2/p (central distributions) are shown. If ΔT^2 is negligible, the variance $\mathrm{V}[T^2/p] \cong 5.06$ from (8). Further, $\mathrm{V}[T^2/p] \cong 5$ under the suppression condition; otherwise, $\mathrm{V}[T^2/p] \gg 5$. This is because the numerical error ΔT^2 affects the sample MD. In Fig. 5(c), the coefficients of variation (C.V.) of the normalized sample MDs T^2/p (central distributions) are shown. If ΔT^2 is negligible, the C.V. $\sqrt{\mathrm{V}[T^2/p]}/\mathrm{E}[T^2/p] \cong 0.96$ from (8). The C.V. shows a similar tendency as the variances in Fig. 5(b). In Fig. 5(d), the slope values of the Q-Q plots for the sample MDs T^2/p (central distributions) are shown. If ΔT^2 is negligible, T^2/p follows an F-distribution, $\mathrm{F}(p, n-p) \cdot (n-1)/(n-p)$. The slope values are nearly 1 and the Q-Q plots nearly coincide with the line $y = x$, which means that the sample MD follows the F-distribution under the suppression condition. Otherwise, the slope values are approximately below 0.8 and the sample MD does not follow the F-distribution. In Fig. 5(e), the minimum misclassification probabilities $(\alpha+\beta)_{min}$ of the sample MDs at non-centrality $\delta = 20$ are shown. Here, $(\alpha+\beta)_{min}$ are the smallest and the sample MDs show the best classification performance under the suppression condition.

As a result of the numerical experiments, the parameters η_1 and η_2, which consist of the maximum and minimum eigenvalues of the covariance matrix and the relative error of the observed sample vector, are smaller than the upper lim-

its given by (5) and (9) with the probability of $1-\alpha$. Hence, the numerical error ΔT^2 of the sample MD T^2 is suppressed and the classification performance of T^2 is not degraded.

4. Conclusion

Thus far, the sample Mahalanobis distance $T^2 = y'\mathrm{S}^{-1}y$ with sample covariance matrix S and observed sample vector y has not been investigated theoretically in terms of the effect of the numerical error ΔT^2 on T^2. To clarify the upper limit of the ratio of ΔT^2 to T^2, the property of ΔT^2 was investigated not by specific case studies but by general theoretical analysis and numerical experiments. As a result, the theoretical conditions for suppressing the effect of ΔT^2 on T^2 were obtained and confirmed for the first time. The theoretical analysis showed that the effect of the numerical error is given not only by the condition number of S and ε_F (the relative error of calculating floating-point real-number variables) but also by the relative error of y. The analysis also showed that as the numerical error increases, the upper limit of the variance of T^2 and the upper limit of the misclassification probability both increase. Thus, if the numerical error ΔT^2 is larger than the statistical variance of T^2 following the F-distribution without the numerical error, T^2 has a broader distribution than the F-distribution and the classification performance of T^2 is degraded. Furthermore, the conditions for suppressing ΔT^2 were proposed as follows. Let the maximum and minimum eigenvalues of S be l_{max} and l_{min}, and let the numerical error of the observed sample vector y be ε. First, $\sqrt{l_{min}/l_{max}} > \varepsilon_F$ needs to be satisfied. Second, the upper limit of $\left|\Delta T^2\right|/T^2$, which consists of $\sqrt{l_{min}/l_{max}}$, ε_F, and ε, needs to be smaller than R_{max}, determined by a tolerance probability α with the F-distribution. In particular, if $\sqrt{l_{min}/l_{max}} \gg \varepsilon_F$, the second condition can be approximated by $\sqrt{l_{min}/l_{max}} > \varepsilon \cdot 2/R_{max}$. Thus, ε (the relative error of the observed sample vector y) governs the numerical error of a sample Mahalanobis distance T^2 if ε is larger than ε_F.

If the above conditions are not satisfied, T^2 has a large numerical error. For example, one way to suppress the error is to regularize S by $\mathrm{S}+\lambda\mathrm{I}$, where I is the unit matrix and the constant λ (the maximum and minimum eigenvalues of $\mathrm{S}+\lambda\mathrm{I}$ are $l_{max}+\lambda$ and $l_{min}+\lambda$, respectively) satisfies $C \gg \varepsilon_F$ and $C > \varepsilon \cdot 2/R_{max}$, where $C = \sqrt{(l_{min}+\lambda)/(l_{max}+\lambda)}$.

The above upper limit of $\left|\Delta T^2\right|/T^2$ clarifies one aspect of ΔT^2 and the conditions for suppressing ΔT^2. However, the stochastic property of ΔT^2, e.g., its probability density function, has not been elucidated thus far. To study the stochastic property, distributions of some elements included by T^2 must be considered, e.g., distributions of sample eigenvalues l_i and the corresponding sample eigenvectors, as well as those of both numerical errors ε_F and ε. However, it is difficult to obtain these distributions. Therefore, elucidating the stochastic property of ΔT^2 theoretically seems to be much more difficult than studying the upper limit of $\left|\Delta T^2\right|/T^2$.

References

[1] J.E. Jackson, A User's Guide to Principal Components, John Wiley & Sons, 2003.
[2] I.T. Jolliffe, Principal Component Analysis 2nd ed., Springer, 2010.
[3] G.E.P. Box, W.G. Hunter, J.F. MacGregor, and J. Erjavac, "Some problems associated with the analysis of multiresponse data," Technometrics, vol.15, pp.33–51, 1973.
[4] J. Bibby, "Some effects of rounding optimal estimates," Sankhyā, vol.42, series B, pts.3 and 4, pp.165–178, 1980.
[5] B.F. Green, Jr., "Parameter sensitivity in multivariate methods," Mult. Behav. Res., vol.12, pp.263–287, 1977.
[6] T. Takeshita, F. Kimura, and K. Miyake, "On the estimation error of Mahalanobis distance," IEICE Trans. Inf & Syst. (Japanese Edition), vol.J70-D, no.3, pp.567–573, 1987.
[7] S. Makinouchi and T. Torii, Numerical Analysis, Ohm, Tokyo, 1975 (in Japanese).
[8] K. Ozawa, "Existence and estimation of the moments of absolute round off error in floating-point arithmetic," Information Processing Society of Japan, vol.22, no.2, pp.139–147, 1981 (in Japanese).
[9] J.H. Friedman, "Regularized Discriminant Analysis," J. American Statistical Association, vol.84, no.405, pp.165–175, March 1989.
[10] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed., pp.53–54, Academic Press, New York, 1990.
[11] D.N. Lawley, "Tests of Significance for the Latent Roots of Covariance and Correlation Matrices," Biometrika, vol.43, pp.128–136, 1956.
[12] H. Nagasaka, "14.3 convergence discrimination and precision of eigenvalues," in Computer and Numerical Analysis, pp.109–117, Asakura-shoten, Tokyo, 1980 (in Japanese).

Appendix:

The upper limit of the relative numerical error $\left|\Delta T^2\right|/T^2$ can be shown as follows.

Let the mean vector $\bar{x}$ be calculated from $x_1, \cdots, x_n$ and define the data matrix $\mathrm{X} \equiv (x_1 - \bar{x}, \cdots, x_n - \bar{x})'$. Let the sample covariance matrix S be $\mathrm{S} = X'X/(n-1)$ with maximum eigenvalue l_{max} and minimum eigenvalue l_{min}. Applying $z \equiv \mathrm{S}^{-1/2}(y-\bar{x})$ to the sample MD for the observed sample vector $T^2 = (y-\bar{x})'\,\mathrm{S}^{-1}\,(y-\bar{x})$, T^2 is expressed as $T^2 = z'z = \|z\|^2$. Consider a matrix A, given by $\mathrm{A} \equiv \mathrm{S}^{1/2}$, where the eigenvalues of A are equal to the singular values of the data matrix X (square roots of the eigenvalues of S). The norms of A and A^{-1} are given by

$$\|\mathrm{A}\| = \sqrt{l_{max}}, \quad \left\|\mathrm{A}^{-1}\right\| = 1/\sqrt{l_{min}}. \tag{A·1}$$

Let the numerical errors of A, y, and z be ΔA, Δy, and Δz respectively. Substituting $y - \bar{x} = \mathrm{A}z$ into $(y-\bar{x}) + \Delta y = (\mathrm{A}+\Delta A)(z+\Delta z)$, Δz is given by

$$\Delta z = \left(\mathrm{I} + \mathrm{A}^{-1}\Delta\mathrm{A}\right)^{-1}\mathrm{A}^{-1}\left\{-(\Delta\mathrm{A})\,z + \Delta y\right\}. \tag{A·2}$$

Supposing that $z \gg \Delta z$ and applying (A·2) to $\Delta T^2 = \|z+\Delta z\|^2 - \|z\|^2 \cong 2z'\Delta z$, ΔT^2 is given by

$$\Delta T^2 = 2z'\left(\mathrm{I} + \mathrm{A}^{-1}\Delta\mathrm{A}\right)^{-1}\mathrm{A}^{-1}\left\{-(\Delta\mathrm{A})\,z + \Delta y\right\}. \tag{A·3}$$

The norms of both sides of (A·3) satisfy (A·4).

$$\left|\Delta T^2\right| \leq 2\|z\| \left\|\left(\mathrm{I}+\mathrm{A}^{-1}\Delta\mathrm{A}\right)^{-1}\right\| \left\|\mathrm{A}^{-1}\right\| \{\|\Delta\mathrm{A}\|\|z\| + \|\Delta y\|\}. \quad (\mathrm{A}\cdot 4)$$

Supposing that $\|\Delta\mathrm{A}\|\left\|\mathrm{A}^{-1}\right\| = \|\Delta\mathrm{A}\| / \sqrt{l_{min}} < 1$ for the norm of $\left\|\left(\mathrm{I}+\mathrm{A}^{-1}\Delta\mathrm{A}\right)^{-1}\right\|$ of (A·4) to exist,

$$\left|\Delta T^2\right| \leq 2\|z\| \frac{\left\|\mathrm{A}^{-1}\right\|}{1-\left\|\mathrm{A}^{-1}\right\|\|\Delta\mathrm{A}\|} \{\|\Delta\mathrm{A}\|\|z\| + \|\Delta y\|\}. \quad (\mathrm{A}\cdot 5)$$

Dividing both sides of (A·5) by $T^2 = \|z\|^2$,

$$\frac{\left|\Delta T^2\right|}{T^2} \leq \frac{2\left\|\mathrm{A}^{-1}\right\|}{1-\left\|\mathrm{A}^{-1}\right\|\|\Delta\mathrm{A}\|}\left\{\|\Delta\mathrm{A}\| + \frac{\|\Delta y\|}{\|z\|}\right\}. \quad (\mathrm{A}\cdot 6)$$

Applying $\|y-\bar{x}\| \leq \|\mathrm{A}\|\|z\|$ to the second term of the right-hand side of (A·6),

$$\frac{2\left\|\mathrm{A}^{-1}\right\|}{1-\left\|\mathrm{A}^{-1}\right\|\|\Delta\mathrm{A}\|} \cdot \frac{\|\Delta y\|}{\|z\|} = \frac{2\left\|\mathrm{A}^{-1}\right\|\|\mathrm{A}\|}{1-\left\|\mathrm{A}^{-1}\right\|\|\Delta\mathrm{A}\|} \cdot \frac{\|\Delta y\|}{\|\mathrm{A}\|\|z\|} \leq \frac{2\left\|\mathrm{A}^{-1}\right\|\|\mathrm{A}\|}{1-\left\|\mathrm{A}^{-1}\right\|\|\Delta\mathrm{A}\|} \cdot \frac{\|\Delta y\|}{\|y-\bar{x}\|}. \quad (\mathrm{A}\cdot 7)$$

Substituting (A·1) and (A·7) into (A·6),

$$\frac{\left|\Delta T^2\right|}{T^2} \leq \frac{2}{1-\dfrac{\|\Delta\mathrm{A}\|}{\sqrt{l_{min}}}}\left\{\frac{\|\Delta\mathrm{A}\|}{\sqrt{l_{min}}} + \sqrt{\frac{l_{max}}{l_{min}}}\cdot\varepsilon\right\}, \quad (\mathrm{A}\cdot 8)$$

where $\varepsilon = \|\Delta y\| / \|y-\bar{x}\|$ denotes the relative numerical error of y, the observed sample vector.

Here, the meaning and the calculation method of $\|\Delta\mathrm{A}\|$ are explained. Since $\Delta\mathrm{A}$ is the matrix of numerical error elements of $\mathrm{A} \equiv \mathrm{S}^{1/2}$, $\|\Delta\mathrm{A}\|$ is the maximum numerical error of the eigenvalues of $\mathrm{S}^{1/2}$. For a symmetric matrix M with true eigenvalues λ_i $(i = 1, \cdots, p)$, if there exists a vector $x \neq 0$ for arbitrary $\epsilon > 0$ that satisfies the sufficient condition of (A·9), the necessary condition of (A·9) holds.

$$\|\mathrm{M}x - \sigma x\| \leq \epsilon\|x\| \rightarrow \min_j \left|\lambda_j - \sigma\right| \leq \epsilon. \quad (\mathrm{A}\cdot 9)$$

From the eigenvalue σ and the eigenvector v with numerical errors, the approximate difference from the true eigenvalue λ_j can be obtained using (A·9). Therefore, the value of $\|\Delta\mathrm{A}\|$, ε_A, is calculated using (A·10) with l_j and v_j, the eigenvalue and eigenvector of A. Furthermore, ε_A can be obtained from (A·11) according to Ref. [12], where ε_F is the relative error of the real variables of T^2 calculation.

$$\varepsilon_A \cong \max_j \frac{\left\|\mathrm{A}v_j - l_j v_j\right\|}{\left\|v_j\right\|}, \quad (\mathrm{A}\cdot 10)$$

$$\varepsilon_A \cong \sqrt{l_{max}}\cdot\varepsilon_F. \quad (\mathrm{A}\cdot 11)$$

Thus, the upper limit of the relative numerical error $\left|\Delta T^2\right| / T^2$ of the sample MD T^2 is given by (A·12), where $\eta_1 = \varepsilon_A / \sqrt{l_{min}} \cong \sqrt{l_{max}/l_{min}}\cdot\varepsilon_F$ and $\eta_2 = \sqrt{l_{max}/l_{min}}\cdot\varepsilon$.

$$\frac{\left|\Delta T^2\right|}{T^2} \leq \frac{2}{1-\dfrac{\varepsilon_A}{\sqrt{l_{min}}}}\left\{\frac{\varepsilon_A}{\sqrt{l_{min}}} + \sqrt{\frac{l_{max}}{l_{min}}}\cdot\varepsilon\right\} = \frac{2(\eta_1+\eta_2)}{1-\eta_1}. \quad (\mathrm{A}\cdot 12)$$

Yasuyuki Kobayashi received his B.E. and M.E. degrees in Electronic Engineering from the University of Tokyo in 1997 and 1999, respectively. From 1999 to 2010, he worked in Mitsubishi Heavy Industries. Since 2011, he has been with the Faculty of Science and Engineering, Teikyo University. He is engaged in research on machine learning and prognostics.