IEICE TRANS. INF. & SYST., VOL.E95–D, NO.12 DECEMBER 2012

LETTER

On d-Asymptotics for High-Dimensional Discriminant Analysis with Different Variance-Covariance Matrices

Takanori AYANO[†*a)], ***Student Member and*** **Joe SUZUKI**[†], ***Member***

SUMMARY In this paper we consider the two-class classification problem with high-dimensional data. It is important to find a class of distributions such that we cannot expect good performance in classification for any classifier. In this paper, when two population variance-covariance matrices are different, we give a reasonable sufficient condition for distributions such that the misclassification rate converges to the worst value as the dimension of data tends to infinity for any classifier. Our results can give guidelines to decide whether or not an experiment is worth performing in many fields such as bioinformatics.

key words: *discriminant analysis, high-dimensional data, misclassification rate, d-asymptotics.*

1. Introduction

Recently in many fields such as microarray analysis, image recognition, and data analysis on the Web we need to analyze high-dimensional data with small sample sizes. For example in microarray analysis the number of data is at most a few hundred, whereas the dimension of data is more than ten thousand. In this paper we study the two-class classification problem with high-dimensional data.

Let $\mathcal{N}_d(\alpha^{(d)}, A^{(d)})$ and $\mathcal{N}_d(\beta^{(d)}, B^{(d)})$ be the d-variate normal distributions with means $\alpha^{(d)}, \beta^{(d)}$ and variance-covariance matrices $A^{(d)}, B^{(d)}$, respectively. We regard $\alpha^{(d)}, \beta^{(d)}, A^{(d)}, B^{(d)}$ as sequences with respect to d. Let x be an observation vector on an individual belonging to $\mathcal{N}_d(\alpha^{(d)}, A^{(d)})$ or to $\mathcal{N}_d(\beta^{(d)}, B^{(d)})$. Our aim is to decide whether x comes from $\mathcal{N}_d(\alpha^{(d)}, A^{(d)})$ or from $\mathcal{N}_d(\beta^{(d)}, B^{(d)})$. A classification rule is defined by a set $G \subset \mathbb{R}^d$. We attribute x to $\mathcal{N}_d(\beta^{(d)}, B^{(d)})$ if $x \in G$ and to $\mathcal{N}_d(\alpha^{(d)}, A^{(d)})$ otherwise. For a classification rule $G \subset \mathbb{R}^d$ the misclassification rate $R(G)$ (with prior probabilities 1/2) is

$$R(G) = \frac{1}{2}\int_G f_{\alpha^{(d)}}(x)\,dx + \frac{1}{2}\int_{\mathbb{R}^d \setminus G} f_{\beta^{(d)}}(x)\,dx,$$

where $f_{\alpha^{(d)}}(x)$ and $f_{\beta^{(d)}}(x)$ are the density functions of $\mathcal{N}_d(\alpha^{(d)}, A^{(d)})$ and $\mathcal{N}_d(\beta^{(d)}, B^{(d)})$, respectively. In practical situations the parameters $\alpha^{(d)}, \beta^{(d)}, A^{(d)}, B^{(d)}$ are unknown. Therefore we need to estimate them from a data set.

The asymptotic property with respect to the dimension of data has been addressed by many authors. For example Fan and Fan [4] gave an upper bound of the misclassification rate for a classifier when both the dimension of data and the number of data tend to infinity while assuming $A^{(d)} = B^{(d)}$. Also Aoshima and Yata [1] proposed a classifier $\hat{G}$ such that the misclassification rate $\mathbf{E}[R(\hat{G})]$ converges to zero as the dimension d tends to infinity under a condition without assuming $A^{(d)} = B^{(d)}$, where $\mathbf{E}$ denotes the expectation with respect to the data.

On the other hand it is important to find a class of distributions such that we cannot expect good performance in classification for any classifier. For example, when we do experiments in bioinformatics, we need many resources such as time, raw materials, and costs. If we know in advance that we cannot expect good performance in classification for any classifier, we can stop doing useless experiments. In theory we want to know a class of distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2 as d tends to infinity for any classifier $\hat{G}$. This is interpreted as the fact that no classification rule is better than a simple random guess. For $A^{(d)} = B^{(d)}$ Ingster et al. [5] gave a necessary and sufficient condition for distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2 as d tends to infinity for any classifier $\hat{G}$. In this paper we consider such a condition for $A^{(d)} \neq B^{(d)}$. For example in hypothesis testing it is known as a difficult problem to test the difference between the means of two normally distributed populations when the variances of the two populations are not assumed to be equal (Behrens-Fisher problem). Also in the two-class classification problem it is not easy to give such a condition for $A^{(d)} \neq B^{(d)}$. Although Matsumoto and Wakaki [6] gave the asymptotic expansion of the misclassification rate under a condition for $A^{(d)} \neq B^{(d)}$, it is hard to find a class of distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2 from [6]. If $\alpha_i^{(d)} \to \beta_i^{(d)}$ and $a_{ij}^{(d)} \to b_{ij}^{(d)}$ as $d \to \infty$ for any i, j, $\mathbf{E}[R(\hat{G})]$ does not necessarily converge to 1/2 for any classifier, where $\alpha^{(d)} = (\alpha_1^{(d)}, \ldots, \alpha_d^{(d)})^T$, $\beta^{(d)} = (\beta_1^{(d)}, \ldots, \beta_d^{(d)})^T$, $A^{(d)} = (a_{ij}^{(d)})$, and $B^{(d)} = (b_{ij}^{(d)})$. In fact, if $\alpha^{(d)} = \beta^{(d)}, A^{(d)} = a^{(d)} I_d, B^{(d)} = b^{(d)} I_d$, and $a^{(d)} \to b^{(d)}$ satisfying $(a^{(d)} - b^{(d)})^4 d \to \infty$, there exists a classifier $\hat{G}$ such that $\mathbf{E}[R(\hat{G})] \to 0$ as $d \to \infty$ (cf. [1] p.372). In this paper, although we do not obtain a necessary and sufficient condition, for $A^{(d)} \neq B^{(d)}$ we give a reasonable sufficient condition for distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2 as d tends to infinity for any classifier $\hat{G}$, i.e., we show that if $\alpha_i^{(d)} \to \beta_i^{(d)}$ and $a_{ij}^{(d)} \to b_{ij}^{(d)}$ fast enough, we cannot expect good performance in classification for any classifier.

Manuscript received July 2, 2012.
Manuscript revised August 28, 2012.

[†]The authors are with the Department of Mathematics, Graduate School of Science, Osaka University, Toyonaka-shi, 560–0043 Japan.

[*]Research fellow of the Japan Society for the Promotion of Science.

a) E-mail: t-ayano@cr.math.sci.osaka-u.ac.jp

DOI: 10.1587/transinf.E95.D.3106

There exists no previous work where a sufficient condition is given for $A^{(d)} \neq B^{(d)}$. Our results can give guidelines to decide whether or not an experiment is worth performing in many fields such as bioinformatics.

Throughout this paper we use the following notations: $\mathbb{R}$ is the set of reals. For $x \in \mathbb{R}^d$ $\|x\|$ denotes the Euclidean norm of x. For a square matrix M, $|M|$ and M^T denote the determinant of M and the transpose of M, respectively. The matrix I_d denotes the unit matrix with order d.

2. Main Result

Since $A^{(d)}$ is symmetric and positive definite, there exists a regular matrix P such that $PA^{(d)}P^T = I_d$. Let $\lambda_1^{(d)}, \ldots, \lambda_d^{(d)}$ be the eigenvalues of $PB^{(d)}P^T$. Since $|PB^{(d)}P^T - \lambda I_d| = |P||B^{(d)} - \lambda A^{(d)}||P^T|$, $\lambda_1^{(d)}, \ldots, \lambda_d^{(d)}$ are the roots of $|B^{(d)} - \lambda A^{(d)}| = 0$. Since $PB^{(d)}P^T$ is symmetric and positive definite, $\lambda_1^{(d)}, \ldots, \lambda_d^{(d)}$ are positive reals. There exists an orthogonal matrix Q such that $QPB^{(d)}P^TQ^T = \Lambda_d$, where $\Lambda_d = \mathrm{diag}(\lambda_1^{(d)}, \ldots, \lambda_d^{(d)})$. Set $R = QP$, then $RA^{(d)}R^T = I_d$ and $RB^{(d)}R^T = \Lambda_d$. Let $\epsilon^{(d)} = \min_{1\le i\le d}\{1 - (\lambda_i^{(d)})^{-1}\}$, $\delta^{(d)} = \max_{1\le i\le d}\{\lambda_i^{(d)} - 1\}$, $\gamma^{(d)} = R(\beta^{(d)} - \alpha^{(d)})$, and $\gamma^{(d)} = (\gamma_1^{(d)}, \ldots, \gamma_d^{(d)})^T$. Then we have the following theorem.

Theorem. *Assume the following conditions:*

[I] For any i and d we have $\lambda_i^{(d)} > 1$, *and there exists* $C_0 > 0$ *such that* $(1/\epsilon^{(d)}) - (1/\delta^{(d)}) \le C_0$ *for any d.*

[II] $\displaystyle\sum_{i=1}^{d}(\lambda_i^{(d)} - 1) \to 0 \quad as \quad d \to \infty.$

[III] $\displaystyle\sum_{i=1}^{d}\left(\frac{\gamma_i^{(d)}}{\lambda_i^{(d)} - 1}\right)^2 \to 0 \quad as \quad d \to \infty.$

Then we have

$$\lim_{d\to\infty} \inf_{\hat{G}} \mathbf{E}[R(\hat{G})] = 1/2,$$

where $\inf_{\hat{G}}$ *denotes the infimum over all the classifiers.*

Proof. For simplicity we omit the suffix (d). Let $G^* := \{x \in \mathbb{R}^d \mid f_\alpha(x) \le f_\beta(x)\}$. Then G^* is the optimal classifier, i.e., $R(G^*) = \inf_{\hat{G}} \mathbf{E}[R(\hat{G})]$, where $\inf_{\hat{G}}$ denotes the infimum over all the classifiers. Therefore it is sufficient to prove that $\lim_{d\to\infty} R(G^*) = 1/2$. From

$$f_\alpha(x) = \left(\frac{1}{2\pi}\right)^{d/2} |A|^{-1/2} \exp\left\{-\frac{1}{2}(x-\alpha)^T A^{-1}(x-\alpha)\right\}$$

and

$$f_\beta(x) = \left(\frac{1}{2\pi}\right)^{d/2} |B|^{-1/2} \exp\left\{-\frac{1}{2}(x-\beta)^T B^{-1}(x-\beta)\right\}$$

we have $G^* = \{x \in \mathbb{R}^d \mid (x-\alpha)^T A^{-1}(x-\alpha) - (x-\beta)^T B^{-1}(x-\beta) + \log|AB^{-1}| \ge 0\}$. Set $y = R(x-\alpha)$, then

$$\begin{aligned} R(G^*) &= \frac{1}{2}\int_{G^*} f_\alpha(x)\,dx + \frac{1}{2}\int_{\mathbb{R}^d\setminus G^*} f_\beta(x)\,dx \\ &= \frac{1}{2}\int_{H_1}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy + \frac{1}{2}\int_{\mathbb{R}^d\setminus H_1} \\ &\quad\cdot\left(\frac{1}{2\pi}\right)^{d/2}|\Lambda_d|^{-1/2}\exp\left\{-\frac{1}{2}(y-\gamma)^T\Lambda_d^{-1}(y-\gamma)\right\}dy, \end{aligned} \tag{1}$$

where

$$H_1 = \left\{y \in \mathbb{R}^d \Big| \sum_{i=1}^{d}(1-\lambda_i^{-1})\left(y_i - \frac{\gamma_i}{1-\lambda_i}\right)^2 \ge \sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i - 1} + \log\lambda_i\right)\right\}.$$

Set $z_i = \lambda_i^{-1/2}(y_i - \gamma_i)$ for (1), then

$$\begin{aligned} R(G^*) &= \frac{1}{2}\int_{H_1}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy \\ &\quad + \frac{1}{2}\int_{H_2}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} z_i^2\right)dz, \end{aligned}$$

where

$$H_2 = \left\{z \in \mathbb{R}^d \Big| \sum_{i=1}^{d}(\lambda_i - 1)\left(z_i - \frac{\lambda_i^{1/2}\gamma_i}{1-\lambda_i}\right)^2 < \sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i - 1} + \log\lambda_i\right)\right\}.$$

Since $\lambda_i > 1$ for any i, we have $\epsilon, \delta > 0$. Therefore

$$\begin{aligned} R(G^*) &\ge \frac{1}{2}\int_{\|y-p\|\ge q}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy \\ &\quad + \frac{1}{2}\int_{\|y-r\|<s}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy, \end{aligned}$$

where

$$p = \left(\frac{\gamma_1}{1-\lambda_1}, \ldots, \frac{\gamma_d}{1-\lambda_d}\right)^T,\ r = \left(\frac{\lambda_1^{1/2}\gamma_1}{1-\lambda_1}, \ldots, \frac{\lambda_d^{1/2}\gamma_d}{1-\lambda_d}\right)^T,$$

$$q = \sqrt{\frac{1}{\epsilon}\sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i - 1} + \log\lambda_i\right)},\ s = \sqrt{\frac{1}{\delta}\sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i - 1} + \log\lambda_i\right)}.$$

Let $\theta = q + \|p\|$ and $\eta = s - \|r\|$, then by the triangle inequality we have $\{y \in \mathbb{R}^d \mid \|y\| \ge \theta\} \subset \{y \in \mathbb{R}^d \mid \|y - p\| \ge q\}$ and $\{y \in \mathbb{R}^d \mid \|y\| < \eta\} \subset \{y \in \mathbb{R}^d \mid \|y - r\| < s\}$. Therefore

$$\begin{aligned} R(G^*) &\ge \frac{1}{2}\int_{\|y\|\ge\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy \\ &\quad + \frac{1}{2}\int_{\|y\|<\eta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy. \end{aligned}$$

Since $\lambda_i > 1$ for any i, we have $\epsilon < \delta$, i.e., $\eta < \theta$. Therefore

$$R(G^*) \ge \frac{1}{2}\left\{1 - \int_{\eta\le\|y\|<\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d} y_i^2\right)dy\right\}. \tag{2}$$

Let $\phi : \mathbb{R}^d \to \mathbb{R}$ be the function defined by $\phi(y) = \|y\|$.

Let μ be the measure on $\mathbb{R}$ induced by ϕ and the Lebesgue measure on $\mathbb{R}^d$. Then by the formula of change of variables (cf. [2], p.216, Theorem 16.13) we have

$$\int_{\eta\le\|y\|<\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\sum_{i=1}^{d}y_i^2\right)dy = \int_{\eta}^{\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\,t^2\right)\mu(dt).$$

Let $F(t)$ be the Lebesgue measure of $\{y\in\mathbb{R}^d \mid \|y\|\le t\}$. Then it is well-known that

$$F(t)=\begin{cases}0 & \text{for } t<0\\ \dfrac{\pi^{d/2}\,t^d}{\Gamma(d/2+1)} & \text{for } t\ge 0,\end{cases}$$

where $\Gamma(\cdot)$ is Gamma function (cf. [3]). Therefore

$$\frac{d\mu}{dt}=\begin{cases}0 & \text{for } t<0\\ \dfrac{d\,\pi^{d/2}\,t^{d-1}}{\Gamma(d/2+1)} & \text{for } t\ge 0,\end{cases}$$

where $(d\mu)/(dt)$ denotes the density function of μ. Therefore

$$\int_{\eta}^{\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\,t^2\right)\mu(dt) = \int_{\max\{\eta,0\}}^{\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\,t^2\right)\frac{d\,\pi^{d/2}}{\Gamma(d/2+1)}\,t^{d-1}\,dt.$$

Let

$$g(t):=\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}\,t^2\right)\frac{d\,\pi^{d/2}}{\Gamma(d/2+1)}\,t^{d-1}.$$

Then $g(t)$ reaches its maximum at $t=\sqrt{d-1}$. By Stirling's formula there exists $C_1>0$ (which does not depend on d) such that

$$g(\sqrt{d-1})=\left(\frac{1}{2\pi}\right)^{d/2}\exp\left\{-\frac{d-1}{2}\right\}\frac{d\,\pi^{d/2}}{\Gamma(d/2+1)}\,(d-1)^{(d-1)/2}$$
$$\le C_1\left(\frac{1}{2\pi}\right)^{d/2}\exp\left\{-\frac{d-1}{2}\right\}\frac{d\,\pi^{d/2}}{d^{\frac{d+1}{2}}\left(\frac{1}{2}\right)^{\frac{d}{2}}\exp(-\frac{d}{2})}(d-1)^{(d-1)/2}$$
$$= C_1\,\exp\left(\frac{1}{2}\right)\left(\frac{d-1}{d}\right)^{(d-1)/2}<\infty.$$

Therefore there exists $C_2>0$ (which does not depend on d) such that $g(t)\le C_2$ for any $t\ge 0$. Therefore

$$\int_{\max\{\eta,0\}}^{\theta}\left(\frac{1}{2\pi}\right)^{d/2}\exp\left(-\frac{1}{2}t^2\right)\frac{d\,\pi^{d/2}}{\Gamma(d/2+1)}t^{d-1}\,dt\le C_2(\theta-\eta).$$

On the other hand we have the following claim.

Claim. $\theta-\eta\to 0$ *as* $d\to\infty$. *(See Appendix for proof.)*

Therefore from (2) and $R(G^*)\le 1/2$ we obtain $\lim_{d\to\infty}R(G^*)=1/2$. The proof is complete. □

3. Conclusion

In this paper for $A^{(d)}\neq B^{(d)}$ we gave a reasonable sufficient condition for distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2 as d tends to infinity for any classifier $\hat{G}$. For $A^{(d)}=B^{(d)}$ Ingster et al. [5] gave a necessary and sufficient condition for distributions such that $\mathbf{E}[R(\hat{G})]$ converges to 1/2, but there exists no previous work where such a condition is given for $A^{(d)}\neq B^{(d)}$. Our results can give guidelines to decide whether or not an experiment is worth performing in many fields such as bioinformatics.

Acknowledgements

This research was supported by Grant-in-Aid for JSPS Fellows (22-2421) from Japan Society for the Promotion of Science.

References

[1] M. Aoshima and K. Yata, "Two-stage procedures for high-dimensional data," Sequential Analysis, vol.30, no.4, pp.356–399, 2011.
[2] P. Billingsley, Probability and Measure, Third Edition, Wiley, 1995.
[3] H.P. Evans, "Volume of an n-dimensional sphere," The American Mathematical Monthly, vol.54, no.10, Part 1, pp.592–594, 1947.
[4] J. Fan and Y. Fan, "High-dimensional classification using features annealed independence rules," Annals of Statistics, vol.36, no.6, pp.2605–2637, 2008.
[5] Y.U. Ingster, C. Pouet, and A.B. Tsybakov, "Classification of sparse high-dimensional vectors," Philosophical Transactions of the Royal Society A, vol.367, pp.4427–4448, 2009.
[6] C. Matsumoto and H. Wakaki, "Asymptotic expansion of the quadratic discriminant function when the dimension and sample sizes are large," Technical Report, no.03-14, Hiroshima Statistical Research Group, Hiroshima University, 2003.
http://www.f-edu.u-fukui.ac.jp/~c-matumo/rireki/QD.pdf

Appendix A: Proof of Claim

From [III] we have $\|p\|\to 0$. From [II] there exists $C_3>0$ (which does not depend on d) such that $\lambda_i\le C_3$ for any i. Therefore from [III] we have $\|r\|\to 0$. From [I] and $(1/\epsilon)>1$ we have

$$q-s=\frac{(1/\epsilon)-(1/\delta)}{\sqrt{1/\epsilon}+\sqrt{1/\delta}}\sqrt{\sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i-1}+\log\lambda_i\right)}$$
$$\le C_0\sqrt{\sum_{i=1}^{d}\left(\frac{\gamma_i^2}{\lambda_i-1}+\log\lambda_i\right)}.$$

From [III] we have

$$\sum_{i=1}^{d}\frac{\gamma_i^2}{\lambda_i-1}=\sum_{i=1}^{d}(\lambda_i-1)\left(\frac{\gamma_i}{\lambda_i-1}\right)^2\le C_3\sum_{i=1}^{d}\left(\frac{\gamma_i}{\lambda_i-1}\right)^2\to 0.$$

From [II] we have $\sum_{i=1}^{d}\log\lambda_i\le\sum_{i=1}^{d}(\lambda_i-1)\to 0$.

Therefore $q-s\to 0$. Therefore we obtain $\theta-\eta\to 0$. □