Journal of Multivariate Analysis 102 (2011) 1361–1373

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Classification of a screened data into one of two normal populations perturbed by a screening scheme

Hea-Jung Kim

Department of Statistics, Dongguk University-Seoul, 100715 Seoul, Republic of Korea

ARTICLE INFO

Article history:
Received 1 July 2010
Available online 17 May 2011

AMS subject classifications:
Primary 62H30
Secondary 60E05

Keywords:
EM algorithm
Probability of misclassification
Screened classification analysis
Screened normal data
Two-sided conditioning normal distribution

ABSTRACT

In normal classification analysis, there may be cases where the population distributions are perturbed by a screening scheme. This paper considers a new classification method for screened data that is obtained from the perturbed normal distributions. Properties of each population distribution is considered and the best region for classifying the screened data is obtained. These developments yield yet another optimal rule for the classification. The rule is studied from several aspects such as a linear approximation, error rates, and estimation of the rule using the EM algorithm. Relationships among these aspects as well as investigation of the rule's performance are also considered. The screened classification ideas are illustrated in detail using numerical examples.

1. Introduction

In a two-group classification analysis, the goal is to take an input vector **Y** and to assign it to one of two discrete classes Π_i with a class level $i = 1, 2$. Researchers can use a variety of methods in a two-group classification analysis. Many of these methods assume, either explicitly or implicitly, the multivariate normal distribution for **Y** given the class membership. Anderson [1], Johnson and Wichern [8], and other researchers provide derivations of the classification regions and construct the optimal rules for assigning future cases to classes on the basis of their measured **Y**. Pardoe et al. [14] and Bishop [4] provide a comprehensive review of the statistical methodology of the classification.

In practice, however, researchers may encounter cases where the classes are screened by an interval $\mathbf{C} = (a, b)$ of an underlying external normal variable $Y_0 \sim N(\mu_0, \sigma_0^2)$, and where the distribution of the input vector **Y** is perturbed by the screening scheme. As an example, consider a case where college admission officers wish to set up an objective criterion (with an input vector **Y**) for admitting students for matriculation; however, the admission officers must first ensure that the students have passed the first screening process. The first screening scheme may be defined by the interval **C** of a criterion variable Y_0 (which includes SAT scores, high-school GPA) so that only students who satisfy $Y_0 \in \mathbf{C}$ can proceed to the admission process. In this case, we encounter a crucial problem for applying the normal classification; given the screening scheme $Y_0 \in \mathbf{C}$, the assumption of the multivariate normal distribution for **Y** is not valid. In fact, each screened class distribution of $\mathbf{X} \overset{d}{=} [\mathbf{Y}|Y_0 \in \mathbf{C}]$ belongs to a family of weighted multivariate normal distributions provided $Cov(\mathbf{Y}, Y_0) \neq \mathbf{0}$.

The distribution of **X**, which has been studied by Kim [9], is as follows. Let $\mathbf{Y}^* \sim N_{p+1}(\boldsymbol{\mu}^*, \Sigma^*)$, where

$$\mathbf{Y}^* = (Y_0, \mathbf{Y}^\top)^\top, \qquad \boldsymbol{\mu}^* = (\mu_0, \boldsymbol{\mu})^\top, \quad \text{and} \quad \Sigma^* = \begin{pmatrix} \sigma_0^2 & \sigma_0\boldsymbol{\delta}^\top \\ \sigma_0\boldsymbol{\delta} & \Sigma \end{pmatrix}. \tag{1.1}$$

E-mail address: kim3hj@dongguk.edu.

0047-259X/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2011.05.003

Then the distribution of the screened normal vector $\mathbf{X} \overset{d}{=} [\mathbf{Y}|Y_0 \in \mathbf{C}]$ is $WTN_p^{(a,b)}(\boldsymbol{\mu}^*, \Sigma^*)$, which is a weighted multivariate two-sided conditioning normal distribution (WTN_p). Suppose that $Z_{(A,B)}$ denotes a doubly truncated $N(0, 1)$ variate with respective upper and lower truncation points B and A, where $A < B$; also suppose that its distribution is written as $TN_{(A,B)}(0, 1)$. According to (15) of [2], a stochastic representation of the WTN_p distribution is

$$\mathbf{X} = \boldsymbol{\mu} + Z_{(v(a),v(b))}\boldsymbol{\delta} + (\Sigma - \boldsymbol{\delta}\boldsymbol{\delta}^\top)^{1/2}\mathbf{Z}, \tag{1.2}$$

where $\mathbf{Z}$ is the $N_p(\mathbf{0}, I_p)$ random vector, and it is independent of $Z_{(v(a),v(b))}$, where $v(a) = (a-\mu_0)/\sigma_0$, and $v(b) = (b-\mu_0)/\sigma_0$. So that $\mathbf{X}$ reduces to $N_p(\boldsymbol{\mu}, \Sigma)$ when $\boldsymbol{\delta} = \mathbf{0}$. Thus (1.2) indicates an intrinsic structure of the WTP_p distributions, and it reveals a type of departure from the multivariate normal law. With respect to the continuous but non-normal input vector $\mathbf{X}$, Lachenbruch et al. [12] note that the performance of the normal classification can be very misleading. This is the problem that motivates our investigation.

In this paper, we introduce a two-group classification method that accounts for the screened classes, $\Pi_i, i = 1, 2$, where the screening is conducted via an interval $\mathbf{C}$ of an underlying external normal variable Y_0. This method is associated with a classification with the skew-normal distributions considered by Azzalini and Capitanio [3] and Reza-Zadkarami and Rowhani [15]; however, as far as we know, no studies have offered a detailed examination of the performance of the screened normal classification analysis (SCA). Interest in the SCA comes from both the theoretical and the applied standpoint. From the theoretical view, the SCA considers another class conditional probability distribution $p(\Pi_i \mid \mathbf{x})$, which is associated with (1.2), in an inference stage. Then this distribution used to derive an optimal classification rule and to study its performance. To this end, Section 2 suggests an optimal classification rule that is induced by the WTN_p population distributions, which contain the classical normal classification rule as a special case. Section 3, approximately computes the total probability of misclassification (TPM) of the SCA, and it proposes some measures based on TPM to evaluate the performance of the SCA; these measures include the screening effect and its robustness. Finally, Section 4 describes the EM algorithm so that we may estimate the unknown parameters of the WTN_p population distributions. Section 5 approaches from the applied viewpoint; it provides numerical illustrations, a new multivariate technique for analyzing a screened data, and broadens the utility of the WTN_p distributions.

2. Screened classification rule

Suppose the joint distributions of $\mathbf{Y}^* = (Y_0, \mathbf{Y}^\top)^\top$ associated with two populations Π_i are $\mathbf{Y}^* \sim N_{p+1}(\boldsymbol{\mu}_i^*, \Sigma_i^*)$, and suppose the populations are screened by an underlying external variable Y_0, where the screening condition is $\{a < Y_0 < b\}$ for $i = 1, 2$. Then the distribution of Π_i is that of $[\mathbf{X} \mid \Pi_i] \overset{d}{=} [\mathbf{Y} \mid \Pi_i, a < Y_0 < b] \sim WTN_p^{(a,b)}(\boldsymbol{\mu}_i^*, \Sigma_i^*)$ for $i = 1, 2$, where $\mathbf{X}$ denote vectors of screened measurements from each population. The classification analysis for the screened populations can be developed in the more general context of WTN_p distributions. For the present study, however, we shall restrict ourselves to a rather simple problem of classification between two screened populations, under the assumptions that they are differ only in the location parameters.

Consider the case of two WTN_p population distributions with an equal scale matrix so that $[\mathbf{X} \mid \Pi_1] \sim WTN_p^{(a,b)}(\boldsymbol{\mu}_1^*, \Sigma^*)$ and $[\mathbf{X} \mid \Pi_2] \sim WTN_p^{(a,b)}(\boldsymbol{\mu}_2^*, \Sigma^*)$, where $\boldsymbol{\mu}_i^* = (\mu_{0i}, \boldsymbol{\mu}_i^\top)^\top$, $i = 1, 2$, and Σ^* is the scale matrix defined in (1.1). Assume that $C(i \mid k)$ denote the cost associated with classifying $\mathbf{x}$ into Π_i when in fact the correct decision should be to classify $\mathbf{x}$ into Π_k, $k = 1, 2$. Then, as a direct consequence of Theorem 6.3.1 of [1], the region of classification into Π_1, R_1, is the set of $\mathbf{x}'s$, $\mathbf{x} \in \mathbb{R}^p$, for which

$$\frac{f(\mathbf{x} \mid \Pi_1)}{f(\mathbf{x} \mid \Pi_2)} \geq \frac{\pi_2 C(1 \mid 2)}{\pi_1 C(2 \mid 1)}, \tag{2.1}$$

where π_i is prior probability of Π_i and

$$f(\mathbf{x} \mid \Pi_i) = \phi_k(\mathbf{x}; \boldsymbol{\mu}_i, \Sigma) \frac{\Phi\left(\xi\ v_i(b) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_i)\right) - \Phi\left(\xi\ v_i(a) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_i)\right)}{\Phi(v_i(b)) - \Phi(v_i(a))} \tag{2.2}$$

by Kim [9]. Here $\phi_p(\cdot; \boldsymbol{\mu}, \Sigma)$ is the pdf of the $N_p(\boldsymbol{\mu}, \Sigma)$ variate, $v_i(a) = (a - \mu_{0i})/\sigma_0$, $v_i(b) = (b - \mu_{0i})/\sigma_0$, $\xi = (1 - \boldsymbol{\delta}^\top \Sigma^{-1} \boldsymbol{\delta})^{-1/2}$, and $\boldsymbol{\lambda}^\top = \xi \boldsymbol{\delta}^\top \Sigma^{-1}$. This yields the best regions of classification that minimizes expected cost of misclassification (ECM) given by

$$R_1 : d(\mathbf{x}) \geq \alpha, \quad \text{and} \quad R_2 : d(\mathbf{x}) < \alpha, \tag{2.3}$$

where $\alpha = \log\{\pi_2 C(1 \mid 2)/(\pi_1 C(2 \mid 1))\}$,

$$d(\mathbf{x}) = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}\mathbf{x} + Q(\mathbf{x}) - \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \Sigma^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2),$$

$$Q(\mathbf{x}) = \log\left\{\frac{\Phi(\xi v_1(b) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_1)) - \Phi(\xi v_1(a) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_1))}{\Phi(\xi v_2(b) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_2)) - \Phi(\xi v_2(a) - \boldsymbol{\lambda}^\top(\mathbf{x} - \boldsymbol{\mu}_2))}\right\} + \log\left\{\frac{\Phi(v_2(b)) - \Phi(v_2(a))}{\Phi(v_1(b)) - \Phi(v_1(a))}\right\}.$$