Available at
www.ElsevierMathematics.com
POWERED BY SCIENCE DIRECT®
Journal of Multivariate Analysis 92 (2005) 174–185
Journal of Multivariate Analysis
http://www.elsevier.com/locate/jmva

A generalized Mahalanobis distance for mixed data

A.R. de Leon[a,1] and K.C. Carrière[a,b,*,2]

[a] *Department of Mathematics & Statistics, University of Calgary, Calgary Alb., Canada T2N 1N4*
[b] *Department of Mathematical & Statistical Sciences, 632 Central Academic Building, University of Alberta, Edmonton Alb., Canada T6G 2G1*

Received 3 July 2002

Abstract

A distance for mixed nominal, ordinal and continuous data is developed by applying the Kullback–Leibler divergence to the general mixed-data model, an extension of the general location model that allows for ordinal variables to be incorporated in the model. The distance obtained can be considered as a generalization of the Mahalanobis distance to data with a mixture of nominal, ordinal and continuous variables. Moreover, it includes as special cases previous Mahalanobis-type distances developed by Bedrick et al. (Biometrics 56 (2000) 394) and Bar-Hen and Daudin (J. Multivariate Anal. 53 (1995) 332). Asymptotic results regarding the maximum likelihood estimator of the distance are discussed. The results of a simulation study on the level and power of the tests are reported and a real-data example illustrates the method.
© 2003 Elsevier Inc. All rights reserved.

AMS 2000 subject classifications: 62E20; 62H12; 62F12

Keywords: Latent variable models; Maximum likelihood; Measurement level; Multivariate normal distribution; Polychoric and polyserial correlations; Probit models

*Corresponding author. Department of Mathematical & Statistical Sciences, 632 Central Academic Building, University of Alberta, Edmonton Alb., Canada T6G 2G1.

E-mail address: kc.carriere@ualberta.ca (K.C. Carrière).

[1] Supported in part by a studentship from the Alberta Heritage Foundation for Medical Research (AHFMR) and a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada.

[2] Funded in part by grants from NSERC and AHFMR. K.C. Carrière is a Senior Scholar with AHFMR.

0047-259X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2003.08.006

1. Introduction

The estimation of a statistical distance between populations arises in many multivariate analysis techniques. In discriminant analysis, for example, the classification rule based on the classical Fisherian linear discriminant function for classifying an observation into one of two or more distinct multivariate normal populations reduces to a comparison of so-called Mahalanobis distances (e.g., [16, p. 31]). This approach is a common one in pattern recognition (e.g., [9]). Whereas distance measures for continuous data are well developed [21], those for mixed discrete and continuous data are less so because of the lack of a standard model for such data.

Krzanowski [10,11] was the first to consider the development of mixed-data distances based on Matusita's distance [17]. Another distance was obtained by Bar-Hen and Daudin [3], who applied the Kullback–Leibler divergence [12, pp. 6–7] to the general location model [19] and derived a distance that specializes to the Mahalanobis distance in the absence of nominal variables. Krusińska [9] proposed a weighted Mahalanobis distance for mixed data as the weighted sum of the Mahalanobis distance for continuous variables and a Mahalanobis-type distance for discrete variables introduced by Kurczyński [13]. More recently, Bedrick et al. [4] derived a Mahalanobis distance for the mixed ordinal and continuous data using the grouped continuous model [2]. However, no distance measure has yet been developed for data with mixed nominal, ordinal and continuous variables. Such a distance must account for not only the different levels of measurement in the variables but also the various types of associations among the variables. The aim of this paper is to develop a statistical distance that can be used for data consisting of a mixture of variable types. Specifically, the problem of generalizing the Mahalanobis distance to data with mixed nominal, ordinal and continuous variables is considered. The approach adopted in the paper unifies previous work on the problem by Bedrick et al. [4] and Bar-Hen and Daudin [3].

To develop the distance, a model for the joint distribution of the mixed variables, called the *general mixed-data model* and first proposed by de Leon and Carriére [5], is described in Section 2. A general distance measure for mixed nominal, ordinal and continuous data is then developed in Section 3 by applying the Kullback–Leibler divergence to the general mixed-data model, and the asymptotic distribution of its maximum likelihood estimate (MLE) is obtained. In addition, a large-sample test of hypothesis concerning two mixed-variate populations is derived in Section 4. The finite-sample performance of the test is investigated via simulations in Section 5. Finally, a real-data example is presented in Section 6 to illustrate the distance measure.

2. General mixed-data model

Let $\mathbf{x} = (X_1, \ldots, X_S)^T$ be the binary representation of $\mathbf{u} = (U_1, \ldots, U_D)^T$, a vector of nominal variables with U_d having s_d possible states ($d = 1, \ldots, D$), so that there

are a total of $S = \prod_{d=1}^{D} s_d$ states for $\mathbf{u}$. Each X_s in $\mathbf{x}$ is defined as either 0 or 1 depending on whether $\mathbf{u}$ falls in state s or not ($\sum_{s=1}^{S} X_s = 1$). By the general location model, the distribution of $\mathbf{x}$ is modelled by a product multinomial distribution $[\mathbf{x}; \boldsymbol{\pi}] = \prod_{s=1}^{S} \pi_s^{\mathbf{x}_{(s)}^T \mathbf{x}}$, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_S)^T$ is the vector of state probabilities ($\sum_{s=1}^{S} \pi_s = 1$), and $\mathbf{x}_{(s)}$ is the vector $\mathbf{x}$ with $X_s = 1$.

Let $\mathbf{y} = (Y_1, \ldots, Y_C)^T$ and $\mathbf{y}^* = (Y_1^*, \ldots, Y_Q^*)^T$ be vectors of continuous and unobservable latent variables, respectively. By the general location model, the conditional distribution $[\mathbf{y}, \mathbf{y}^* \mid \mathbf{x}_{(s)}]$ is modelled as multivariate normal with mean $\boldsymbol{\eta}_s$ and common covariance matrix $\boldsymbol{\Gamma}$, given $\mathbf{x} = \mathbf{x}_{(s)}$, where $\boldsymbol{\eta}_s$ and $\boldsymbol{\Gamma}$ are partitioned accordingly as

$$\boldsymbol{\eta}_s = \begin{pmatrix} \boldsymbol{\mu}_s \\ \boldsymbol{\mu}_s^* \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Gamma} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*} \\ \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^T & \boldsymbol{\Sigma}^* \end{pmatrix}. \tag{1}$$

The $CS \times 1$ stacked vector of state means of $\mathbf{y}$ is denoted as $\boldsymbol{\mu}$. The latent relationship between $\mathbf{y}^*$ and the vector of ordinal variables $\mathbf{z} = (Z_1, \ldots, Z_Q)^T$ is defined by the threshold model by which $Z_q = a_q^{\ell_q}$ if and only if $\alpha_q^{\ell_q - 1} < Y_q^* \leqslant \alpha_q^{\ell_q}$, where $\{\alpha_q^0 = -\infty, \alpha_q^1, \ldots, \alpha_q^{L_q}, \alpha_q^{L_q+1} = +\infty\}$ are the unknown cutpoints or thresholds, and $a_q^1 < a_q^2 < \cdots < a_q^{L_q+1}$ are the ordinal scores for Z_q, $q = 1, \ldots, Q$. Note that the set of thresholds as well as the scores vary for each ordinal variable in $\mathbf{z}$ but is constant across states. Without loss of generality, it is assumed that $a_q^{\ell_q} = \ell_q$, $\ell_q = 1, \ldots, L_q + 1$.

Then, under the general location model, the conditional distribution $[\mathbf{y}^* \mid \mathbf{x}_{(s)}, \mathbf{y}]$ is multivariate normal with mean $\boldsymbol{\mu}_s^* + \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^T \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}_s)$ and covariance matrix $\boldsymbol{\Sigma}^* - \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma}_{\mathbf{y}\mathbf{y}^*} \equiv \mathbf{DRD}$, where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_Q)$ is the diagonal matrix of conditional standard deviations and $\mathbf{R} = (r_{qq'})$ is the symmetric matrix of conditional polychoric correlations [6] of $\mathbf{z}$, given $\mathbf{x}_{(s)}$ and $\mathbf{y}$. Similar to the usual development of latent variable models, it may be assumed without loss of generality that $\boldsymbol{\Sigma}^* = \mathbf{R}^*$, the correlation matrix of $\mathbf{y}^*$. To avoid over-parameterizing the model, state S is fixed as a *reference state* and $\boldsymbol{\mu}_s$ and $\boldsymbol{\mu}_s^*$ ($s \neq S$) are defined as $\boldsymbol{\mu}_s = \boldsymbol{\xi} + \boldsymbol{\xi}_s$ and $\boldsymbol{\mu}_s^* = \boldsymbol{\xi}^* + \boldsymbol{\xi}_s^*$, where $\boldsymbol{\xi} = \boldsymbol{\mu}_S$ and $\boldsymbol{\xi}^* = \boldsymbol{\mu}_S^*$, the means of $\mathbf{y}$ and $\mathbf{y}^*$, respectively, and $\boldsymbol{\xi}_s$ and $\boldsymbol{\xi}_s^*$ are the effects of state $s = 1, \ldots, S-1$, relative to that of state S.

Let $\boldsymbol{\ell} = (\ell_1, \ldots, \ell_Q)^T$ be a possible value of $\mathbf{z}$. By a suitable transformation, it can be shown that

$$[\mathbf{z} = \boldsymbol{\ell} \mid \mathbf{x} = \mathbf{x}_{(s)}, \mathbf{y}] = \int_{\mathscr{S}_{I(s,\mathbf{y})}} \phi_Q(\mathbf{v} \mid \mathbf{R})\, \mathrm{d}\mathbf{v}, \tag{2}$$

where $\phi_Q(\cdot \mid \mathbf{R})$ is the Q-dimensional normal density with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, and $\mathscr{S}_{I(s,\mathbf{y})} = \{(v_1, \ldots, v_Q) : v_{sq}^{\ell_q - 1} < v_q \leqslant v_{sq}^{\ell_q}, q = 1, \ldots, Q\}$, with $v_{sq}^{\ell_q} = \gamma_q^{\ell_q} - \tau_{sq} - \boldsymbol{\beta}_q^T \mathbf{y}$. Here, $\gamma_q^{\ell_q} = \alpha_q^{\ell_q}/d_q - (\xi_q^*/d_q - \boldsymbol{\beta}_q^T \boldsymbol{\xi})$, τ_{sq} is the qth element of $\boldsymbol{\tau}_s = \mathbf{D}^{-1} \boldsymbol{\xi}_s^* - \mathbf{B} \boldsymbol{\xi}_s$,

and $\boldsymbol{\beta}_q^T$ is the qth row of $\mathbf{B} = \mathbf{D}^{-1}\boldsymbol{\Sigma}_{\mathbf{yy}^*}^T\boldsymbol{\Sigma}^{-1}$, $\ell_q = 1, \ldots, L_q$. Note that $\tau_{Sq} \equiv 0\ \forall q$, and the extreme cutpoints are taken as $\gamma_q^0 = -\infty$ and $\gamma_q^{L_q+1} = +\infty$.

The joint density $[\mathbf{x}, \mathbf{y}, \mathbf{z}]$ of $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ can thus be written as

$$[\mathbf{x} = \mathbf{x}_{(s)}, \mathbf{y}, \mathbf{z} = \boldsymbol{\ell}] = \pi_s \times \phi_C(\mathbf{y} - \boldsymbol{\mu}_s \mid \boldsymbol{\Sigma}) \int_{\mathscr{S}_{I(s,\mathbf{y})}} \phi_Q(\mathbf{v} \mid \mathbf{R})\, d\mathbf{v}. \tag{3}$$

This joint density is called the *general mixed-data model* with parameter $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_1^T, \boldsymbol{\theta}_2^T, \boldsymbol{\theta}_3^T)$, where $\boldsymbol{\theta}_1^T = (\pi_1, \ldots, \pi_{S-1}), \boldsymbol{\theta}_2^T = (\boldsymbol{\mu}^T, \{\mathrm{vech}(\boldsymbol{\Sigma})\}^T), \boldsymbol{\theta}_3^T = (\boldsymbol{\gamma}^T, \{\mathrm{vech}(\mathbf{R})\}^T, \boldsymbol{\beta}^T, \boldsymbol{\tau}^T)$, with $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_Q^T)$, $\boldsymbol{\tau}^T = (\boldsymbol{\tau}_1^T, \ldots, \boldsymbol{\tau}_{S-1}^T)$, and $\mathrm{vech}(\boldsymbol{\Sigma})$ and $\mathrm{vech}(\mathbf{R})$ are the vectors containing the upper diagonal elements of $\boldsymbol{\Sigma}$ and $\mathbf{R}$, respectively.

The general location model is obtained from the general mixed-data model by setting $Q = 0$, and hence, the former may be viewed as a special case of the latter. Similarly, the general mixed-data model reduces to the conditional grouped continuous model [2] when $S = 1$. Therefore, the general mixed-data model unifies these two mixed-data models into a single model.

2.1. Maximum likelihood estimation

Given a mixed-variable random sample $(\mathbf{x}_i^T, \mathbf{y}_i^T, \mathbf{z}_i^T)^T$, $i = 1, \ldots, N$, define the sets $\mathscr{A}(s) = \{i \mid \mathbf{x}_i = \mathbf{x}_{(s)}\}$ and $\mathscr{B}(\ell_1, \ldots, \ell_Q) = \{i \mid Z_{iq} = \ell_q, \ell_q = 1, \ldots, L_q + 1; q = 1, \ldots, Q\}$. Using (3), the likelihood function can be written as

$$\mathscr{L} = \left[(1 - \pi_1 - \cdots - \pi_{S-1})^{n_S} \prod_{s=1}^{S-1} \pi_s^{n_s}\right] \phi_C^N(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \boldsymbol{\theta}_2) \times \prod_{s=1}^{S} \prod_{\ell_1=1}^{L_1+1} \cdots \prod_{\ell_Q=1}^{L_Q+1} \prod_{i(s,\boldsymbol{\ell})} \sum_{\varepsilon_1=0}^{1} \cdots \sum_{\varepsilon_Q=0}^{1} (-1)^{\sum_{q=1}^{Q} \varepsilon_q + Q} \Phi_Q(\ldots, v_{i(s,\boldsymbol{\ell})q}^{\ell_q - \varepsilon_q}, \ldots \mid \mathbf{R}),$$

where $\phi_C^N(\mathbf{y}_1, \ldots, \mathbf{y}_N \mid \boldsymbol{\theta}_2)$ is the usual multivariate normal likelihood, $\Phi_Q(\cdot \mid \mathbf{R})$ is the Q-dimensional normal distribution function with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$, and $v_{i(s,\boldsymbol{\ell})q}^{\ell_q} = \gamma_q^{\ell_q} - \tau_{sq} - \boldsymbol{\beta}_q^T \mathbf{y}_{i(s,\boldsymbol{\ell})}$. The index "$i(s,\boldsymbol{\ell})$" comes from $\mathscr{A}(s) \cap \mathscr{B}(\ell_1, \ldots, \ell_Q)$, and refers to the ith unit in state s such that $Z_1 = \ell_1, \ldots, Z_Q = \ell_Q$. Note that $v_{i(s,\boldsymbol{\ell})q}^0 = -\infty$ and $v_{i(s,\boldsymbol{\ell})q}^{L_q+1} = +\infty$, $q = 1, \ldots, Q$.

The usual MLE for a multinomial model given by $\hat{\pi}_s = n_s/N$ is obtained as $\hat{\boldsymbol{\theta}}_1$ while $\hat{\boldsymbol{\theta}}_2$ is the usual MLE for a multivariate normal sample which consists of $\hat{\boldsymbol{\mu}}_s = \overline{\mathbf{y}}_s = \sum_{i(s)} \mathbf{y}_{i(s)}/n_s$ and the unique elements of $\hat{\boldsymbol{\Sigma}} = \mathbf{S} = \sum_{s=1}^{S} \sum_{i(s)} (\mathbf{y}_{i(s)} - \overline{\mathbf{y}}_s)(\mathbf{y}_{i(s)} - \overline{\mathbf{y}}_s)^T/N$, where $\overline{\mathbf{y}}_s$ is the sth state mean, $s = 1, \ldots, S$. The MLE $\hat{\boldsymbol{\theta}}_3$ can be obtained via iterative techniques such as the *Fletcher–Powell* algorithm. Details for implementing this are found in de Leon and Carriére [5].

Using standard large-sample results on MLE, it can be easily shown that $\hat{\boldsymbol{\theta}}$ is consistent for $\boldsymbol{\theta}$ and satisfies $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} \xrightarrow{\mathscr{L}} \mathscr{N}_P(\mathbf{0}, \frac{1}{N}\mathscr{I}_P^{-1}(\boldsymbol{\theta}))$ as $N \to \infty$, where $\mathscr{I}_P(\boldsymbol{\theta})$ is the usual expected Fisher information matrix based on all the observations. Large-sample standard errors for the MLEs are obtained from the diagonals of $\mathscr{I}_P^{-1}(\hat{\boldsymbol{\theta}})/N$ (see de Leon and Carriére [5] for more details).

3. A generalized Mahalanobis distance

In this section, a distance is derived for mixed nominal, ordinal and continuous data as modelled by the general mixed-data model described in Section 2. The distance includes as special cases previous generalizations of the Mahalanobis distance to mixed data proposed by Bedrick et al. [4] and Bar-Hen and Daudin [3].

Suppose $(\mathbf{x}_g^T, \mathbf{y}_g^T, \mathbf{z}_g^T)^T$ is a random vector from the mixed-variate population $\mathscr{P}^{(g)}$ defined by the general mixed-data model with parameter $\boldsymbol{\theta}_g$ containing $\boldsymbol{\pi}_g$, $\boldsymbol{\mu}_g$ and $\boldsymbol{\tau}_g$, for $g = 1, \ldots, G$, and $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$, vech($\boldsymbol{\Sigma}$), and vech($\mathbf{R}$). Note that this implies that the populations differ only in their locations. As well, it is assumed that the reference states in each of the populations are the same with $\boldsymbol{\xi}_g^* = \boldsymbol{\xi}^*$ and $\boldsymbol{\xi}_g = \boldsymbol{\xi}$ $\forall g$. This approach is similar to that adopted earlier by Poon and Lee [20] and Lee et al. [15].

The following formal definition of the Kullback–Leibler divergence given by Kullback [12, p. 6] is presented for later use.

Definition 3.1. Let ψ_g, $\psi_{g'}$ and λ be three probability measures absolutely continuous with respect to each other, and assume there exist generalized probability densities f_g and $f_{g'}$, the respective Radon–Nikodym derivatives of ψ_g and $\psi_{g'}$ with respect to λ. The divergence measure between f_g and $f_{g'}$ defined as

$$\Delta_{gg'} = \int [f_g(\mathbf{w}) - f_{g'}(\mathbf{w})] \log \frac{f_g(\mathbf{w})}{f_{g'}(\mathbf{w})} d\lambda,$$

is called the Kullback–Leibler divergence.

Here, $\Delta_{gg'}$ possesses all the properties of a distance except for the triangle inequality, and is therefore not considered a distance [12, Chapter 2].

When f_g is $\mathscr{N}(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$ and $f_{g'}$ is $\mathscr{N}(\boldsymbol{\mu}_{g'}, \boldsymbol{\Sigma})$, then $\Delta_{gg'} = (\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_g - \boldsymbol{\mu}_{g'})$, the Mahalanobis distance between two multivariate normal populations. In this respect, $\Delta_{gg'}$ can be considered as a generalization of the Mahalanobis distance. Bar-Hen and Daudin [3] used $\Delta_{gg'}$ to generalize the Mahalanobis distance to mixed binary and continuous data modelled by the general location model, and derived the asymptotic distribution of its MLE.

Theorem 3.1 below is obtained by applying Definition 3.1 to the general mixed-data models for $\mathscr{P}^{(g)}$ and $\mathscr{P}^{(g')}$.

Theorem 3.1. *The Kullback–Leibler divergence between* $\mathscr{P}^{(g)}$ *and* $\mathscr{P}^{(g')}$ *is given by*

$$\Delta_{gg'} = \Delta^1_{gg'} + \Delta^2_{gg'} + \Delta^3_{gg'}, \tag{4}$$

where

$$\Delta^1_{gg'} = \sum_{s=1}^{S} (\pi_{gs} - \pi_{g's}) \log \frac{\pi_{gs}}{\pi_{g's}},$$

$$\Delta^2_{gg'} = \sum_{s=1}^{S} \frac{\pi_{gs} + \pi_{g's}}{2} (\boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's}),$$

$$\Delta^3_{gg'} = \sum_{s=1}^{S-1} \frac{\pi_{gs} + \pi_{g's}}{2} (\boldsymbol{\tau}_{gs} - \boldsymbol{\tau}_{g's})^T \mathbf{R}^{-1} (\boldsymbol{\tau}_{gs} - \boldsymbol{\tau}_{g's}).$$

Proof. Suppose $\mathbf{y}^*_g$ is the latent variable underlying $\mathbf{z}_g$, and that $(\mathbf{x}^T_g, \mathbf{y}^T_g, \mathbf{y}^{*T}_g)$ follows the general location model with parameters $\boldsymbol{\pi}_g$, $(\boldsymbol{\mu}^T_g, \boldsymbol{\mu}^{*T}_g)^T$, and $\boldsymbol{\Gamma}$, where $\boldsymbol{\mu}^{*T}_g = (\boldsymbol{\mu}^{*T}_{g1}, \ldots, \boldsymbol{\mu}^{*T}_{gS})$ is the $QS \times 1$ stacked vector of state means of $\mathbf{y}^*_g$ and $\boldsymbol{\Gamma}$ is as defined in (1). Using results in Kullback [12, Chapter 6] and Proposition 2.1 in Bar-Hen and Daudin [3], it follows that

$$\Delta_{gg'} = \sum_{s=1}^{S} (\pi_{gs} - \pi_{g's}) \log \frac{\pi_{gs}}{\pi_{g's}} + \sum_{s=1}^{S} \frac{\pi_{gs} + \pi_{g's}}{2} \begin{pmatrix} \boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's} \\ \boldsymbol{\mu}^*_{gs} - \boldsymbol{\mu}^*_{g's} \end{pmatrix}^T \boldsymbol{\Gamma}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's} \\ \boldsymbol{\mu}^*_{gs} - \boldsymbol{\mu}^*_{g's} \end{pmatrix}.$$

By the decomposition of the Mahalanobis distance [16, pp. 78–79],

$$\begin{pmatrix} \boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's} \\ \boldsymbol{\mu}^*_{gs} - \boldsymbol{\mu}^*_{g's} \end{pmatrix}^T \boldsymbol{\Gamma}^{-1} \begin{pmatrix} \boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's} \\ \boldsymbol{\mu}^*_{gs} - \boldsymbol{\mu}^*_{g's} \end{pmatrix} = \left(\boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu}_{gs} - \boldsymbol{\mu}_{g's}\right) + \left(\boldsymbol{\mu}_{gs\cdot y} - \boldsymbol{\mu}_{g's\cdot y}\right)^T (\mathbf{DRD})^{-1} \times \left(\boldsymbol{\mu}_{gs\cdot y} - \boldsymbol{\mu}_{g's\cdot y}\right),$$

where $\boldsymbol{\mu}_{gs\cdot y} = \boldsymbol{\mu}^*_{gs} - \mathbf{DB}\boldsymbol{\mu}_{gs}$, with $\mathbf{D}$ and $\mathbf{B}$ as defined in Section 2, $g = g', g''$. Since $\boldsymbol{\mu}^*_{gs} = \boldsymbol{\xi}^* + \boldsymbol{\xi}^*_{gs}$ and $\boldsymbol{\mu}_{gs} = \boldsymbol{\xi} + \boldsymbol{\xi}_{gs}$ for $s \neq S$, it follows that $\boldsymbol{\mu}_{g's\cdot y} - \boldsymbol{\mu}_{g''s\cdot y} = \boldsymbol{\xi}^*_{g's} - \mathbf{DB}\boldsymbol{\xi}_{g's} - (\boldsymbol{\xi}^*_{g''s} - \mathbf{DB}\boldsymbol{\xi}_{g''s})$. Expression (4) is now immediate by noting from Section 2 that $\boldsymbol{\tau}_{gs} = \mathbf{D}^{-1}\boldsymbol{\xi}^*_{gs} - \mathbf{B}\boldsymbol{\xi}_{gs}$ for $g = g', g''$. □

Remark 3.1. With $Q = 0$, $\Delta_{gg'} = \Delta^1_{gg'} + \Delta^2_{gg'}$ is the distance proposed by Bar-Hen and Daudin [3] while with $S = 1$, $\Delta_{gg'} = \Delta^2_{gg'} + \Delta^3_{gg'}$ corresponds to that by Bedrick et al. [4]. Thus, Theorem 3.1 generalizes these two previous Mahalanobis-type distances for mixed data.

Remark 3.2. $\Delta_{gg'}$ can be considered an extension of the Mahalanobis distance since it reduces to it for $Q = 0, S = 1$. Note also that $\Delta_{gg'} = \Delta_{g'g}$ for any g, g'.

Remark 3.3. Note that when the nominal variables are independent of the continuous and ordinal variables, $\Delta_{gg'}$ is simply the sum of the distances corresponding to each variable type.

Given random samples $(\mathbf{x}_{gi}^T, \mathbf{y}_{gi}^T, \mathbf{z}_{gi}^T)^T$, $i = 1, \ldots, n_{g\cdot}, g = 1, \ldots, G$, the MLE of $\Delta_{gg'}$ is given by $\hat{\Delta}_{gg'} = \hat{\Delta}^1_{gg'} + \hat{\Delta}^2_{gg'} + \hat{\Delta}^3_{gg'}$, where

$$\hat{\Delta}^1_{gg'} = \sum_{s=1}^{S} (\hat{\pi}_{gs} - \hat{\pi}_{g's}) \log \frac{\hat{\pi}_{gs}}{\hat{\pi}_{g's}},$$

$$\hat{\Delta}^2_{gg'} = \sum_{s=1}^{S} \frac{\hat{\pi}_{gs} + \hat{\pi}_{g's}}{2} (\hat{\boldsymbol{\mu}}_{gs} - \hat{\boldsymbol{\mu}}_{g's})^T \hat{\boldsymbol{\Sigma}}^{-1} (\hat{\boldsymbol{\mu}}_{gs} - \hat{\boldsymbol{\mu}}_{g's}),$$

$$\hat{\Delta}^3_{gg'} = \sum_{s=1}^{S-1} \frac{\hat{\pi}_{gs} + \hat{\pi}_{g's}}{2} (\hat{\boldsymbol{\tau}}_{gs} - \hat{\boldsymbol{\tau}}_{g's})^T \hat{\mathbf{R}}^{-1} (\hat{\boldsymbol{\tau}}_{gs} - \hat{\boldsymbol{\tau}}_{g's}),$$

with the unknown parameters simply replaced by their MLEs. The asymptotic distribution of $\hat{\Delta}_{gg'}$ under the hypothesis that $\boldsymbol{\theta}_g = \boldsymbol{\theta}_{g'}$ is derived in the following section.

4. Asymptotic results

Consider the problem of constructing a statistical test of

$$H : \boldsymbol{\theta}_g = \boldsymbol{\theta}_{g'} \quad \text{against } K : \boldsymbol{\theta}_g \neq \boldsymbol{\theta}_{g'}. \tag{5}$$

The following theorem derives a large-sample test of (5). Note that H is equivalent to $H' : \Delta_{gg'} = 0$.

Theorem 4.1. *Suppose $\mathscr{P}^{(g)}$ and $\mathscr{P}^{(g')}$ are mixed-variate populations defined by the general mixed-data models with respective parameters $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_{g'}$. Under $H : \boldsymbol{\theta}_g = \boldsymbol{\theta}_{g'}$, then*

$$\frac{n_{g\cdot} n_{g'\cdot}}{n_{g\cdot} + n_{g'\cdot}} \hat{\Delta}_{gg'} \xrightarrow{\mathscr{L}} \chi^2_P, \tag{6}$$

when $\frac{n_{g\cdot}}{n_{g'\cdot}} \to \delta$ as $n_{g\cdot} \to \infty$, $n_{g'\cdot} \to \infty$, where $\delta < \infty$ and P is the total number of unknown parameters.

Proof. The proof is similar to that of Proposition 3.1 of Bar-Hen and Daudin [3]. Let $\boldsymbol{\theta}$ be the common value of $\boldsymbol{\theta}_g$ and $\boldsymbol{\theta}_{g'}$ under H. Similar to Bar-Hen and Daudin [3], a first-order Taylor series expansion of $\hat{\Delta}_{gg'}$ at a neighborhood

of $(\boldsymbol{\theta}_g, \boldsymbol{\theta}_{g'})$ yields

$$\begin{aligned}\hat{\Delta}_{gg'} &= \Delta_{gg'} + \sum_{g,g'} (\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)^T \frac{\partial \Delta_{gg'}}{\partial \boldsymbol{\theta}_g} + \frac{1}{2} \sum_{g,g'} (\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)^T \frac{\partial^2 \Delta_{gg'}}{\partial \boldsymbol{\theta}_g \partial \boldsymbol{\theta}_g^T} (\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g) \\ &\quad + (\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g)^T \frac{\partial^2 \Delta_{gg'}}{\partial \boldsymbol{\theta}_g \partial \boldsymbol{\theta}_{g'}^T} (\hat{\boldsymbol{\theta}}_{g'} - \boldsymbol{\theta}_{g'}) + \sum_{g,g'} o(\,||\,\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\,||\,) \\ &= (\hat{\boldsymbol{\theta}}_g - \hat{\boldsymbol{\theta}}_{g'})^T \mathcal{I}_P(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_g - \hat{\boldsymbol{\theta}}_{g'}),\end{aligned}$$

under H, where $o(\,||\,\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}_g\,||\,) \xrightarrow{p} 0$ as $\hat{\boldsymbol{\theta}}_g \to \boldsymbol{\theta}_g$ for $g = g, g'$.

From Section 2, it follows that $\sqrt{\frac{n_{g\cdot} n_{g'\cdot}}{n_{g\cdot} + n_{g'\cdot}}}(\hat{\boldsymbol{\theta}}_g - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \frac{1}{1+\delta} \mathcal{I}_P^{-1}(\boldsymbol{\theta}))$ and $\sqrt{\frac{n_{g\cdot} n_{g'\cdot}}{n_{g\cdot} + n_{g'\cdot}}}(\hat{\boldsymbol{\theta}}_{g'} - \boldsymbol{\theta}) \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \frac{\delta}{1+\delta} \mathcal{I}_P^{-1}(\boldsymbol{\theta}))$. Hence, $\sqrt{\frac{n_{g\cdot} n_{g'\cdot}}{n_{g\cdot} + n_{g'\cdot}}} \mathcal{I}_P^{1/2}(\hat{\boldsymbol{\theta}}_g - \hat{\boldsymbol{\theta}}_{g'}) \xrightarrow{\mathcal{L}} \mathcal{N}_P(\mathbf{0}, \mathbf{I}_P)$, and the result follows immediately. □

Remark 4.1. Theorem 4.1 generalizes Proposition 3.1 of Bar-Hen and Daudin [3] to the general mixed-data model. In fact, Proposition 3.1 is obtained by taking $Q = 0$ in Theorem 4.1.

Remark 4.2. A two-sample test for mixed data distributed according to the conditional grouped continuous model is obtained from Theorem 4.1 by taking $S = 1$. Similar tests based on likelihood ratio and generalized Wald statistics are discussed by Lapidus [14, Chapter 4] and by Afifi and Elashoff [1].

The level and power of the test described in Theorem 4.1 are evaluated through simulations in the next section.

5. Simulation study

In the simulations, general mixed-data models with $C = L = Q = 1$ and $S = 2$ are considered. The parameter is then $\boldsymbol{\theta}_g^T = (\pi_g, \boldsymbol{\mu}_g^T, \sigma^2, \gamma, \beta, \tau_g)$, where $\boldsymbol{\mu}_g^T = (\mu_{g1}, \mu_{g2})$ with μ_{gs} the sth state mean of Y_g, γ is the standardized cutpoint α for the latent variable Y_g^* underlying Z_g, and τ_g is the effect of state 1 on Z_g relative to that of state 2. Note that $\gamma = \alpha/\sqrt{1-\rho^2} - (\mu_2^*/\sqrt{1-\rho^2} - \beta\mu_2)$, $\beta = \rho/(\sigma\sqrt{1-\rho^2})$, and $\tau_g = \xi_g^*/\sqrt{1-\rho^2} - \beta\xi_g$, where $\xi_g = \mu_{g1} - \mu_2$, $\xi_g^* = \mu_{g1}^* - \mu_2^*$, with $\mu_2 = \mu_{g2}$, $\mu_2^* = \mu_{g2}^*$, and $\mu_{gs}^* = E(Y_g^* \mid \mathbf{x}_g = \mathbf{x}_{(s)})$, for $g = 1, 2$, and $s = 1, 2$. Note also that $Z_g = 2$ if $Y_g^* > \alpha$ and $Z_g = 1$ if $Y_g^* \leqslant \alpha$. Similar to Bar-Hen and Daudin [3], the following five cases are considered:

(0) no differences between populations with respect to all three variable types;
(a) there is difference between populations only with respect to nominal vector $\mathbf{x}$;
(b) there is difference between populations only with respect to continuous variable Y;

(c) there is difference between populations only with respect to ordinal variable Z;
(d) populations are different with respect to all three variable types.

To assess the size and power of the χ^2 test in Theorem 4.1, random samples of various sizes $(n_1, n_2) = (50, 25), (50, 100), (100, 100)$, and $(100, 150)$ were generated from the general mixed-data models with $(\sigma^2, \rho, \alpha)^T = (1, 0.5, 1)^T$ and $(p_g, \mu_{g1}, \mu_{g2}, \mu^*_{g1}, \mu^*_{g2})^T$, $g = 1, 2$, given by (0) $(0.5, 0, 0.5, 0, 0.5)^T$ for both populations, (a) $(0.5, 0, 0.5, 0, 0.5)^T$ for population 1 and $(0.75, 0, 0.5, 0, 0.5)^T$ for population 2, (b) $(0.5, 0, 0.5, 0, 0.5)^T$ for population 1 and $(0.5, 0.5, 0.5, 0, 0.5)^T$ for population 2, (c) $(0.5, 0, 0.5, 0, 0.5)^T$ for population 1 and $(0.5, 0, 0.5, 0.5, 0.5)^T$ for population 2, and (d) $(0.5, 0, 0.5, 0, 0.5)^T$ for population 1 and $(0.75, 0.5, 0.5, 0.5, 0.5)^T$ for population 2.

Observe that case (0) is taken as having the true parameter configurations for both populations under the null hypothesis $H : \Delta_{12} = 0$. For each combination of case and (n_1, n_2) above, 1000 replications were generated in S-PLUS. Hypothesis H is then rejected if and only if $n_1 n_2 \hat{\Delta}_{12}/(n_1 + n_2) > \chi^2_{7,05} = 14.1$, the 95th percentile of the χ^2 distribution with seven degrees of freedom. Results of the simulated levels and powers of the test are displayed in Table 1.

Three observations are apparent from the table. First, the power of the test increases with the total sample size $n_1 + n_2$. Second, the test tends to be liberal when the total sample size is small, confirming an earlier finding reported by Bar-Hen and Daudin [3]. However, given large enough samples, the test is able to attain the nominal level. Finally, as was similarly reported by Bar-Hen and Daudin [3], the power of the test is higher when differences exist with respect to all three variables than when the difference is only with respect to just one variable.

6. Example

In this section, real data are used to illustrate the distance developed in this paper. The data come from Koepsel et al. [8] (also in [7, pp. 680–683]) and concern the occurrence and non-occurrence of perforation of the appendix. Data from a total of 181 surgery patients are included in the analysis, and four variables are considered. The same data were analyzed by Nakanishi [18] in the context of variable selection in mixed-data discriminant analysis.

For the purpose of this example, variable X_3 as defined in Fisher and Van Bell [7, p. 680] is transformed into an ordinal variable Z with 2 levels (long or short duration). The states of $\mathbf{x}^T = (X_1, X_2)$ correspond with the patient's perforation status, with $\mathbf{x} = \mathbf{x}_{(2)}$ if perforation is present and $\mathbf{x} = \mathbf{x}_{(1)}$ otherwise. The following variables are included:

$Y :=$ time in hours from physician contact to surgery,

$Z :=$ duration in hours of symptoms prior to physician contact

$$= \begin{cases} 2 & \text{no. of hours } > 24, \\ 1 & \text{otherwise.} \end{cases}$$

Table 1
Empirical size and power of χ^2 test in Theorem 4.1 for $C = L = Q = 1$ and $S = 2$ based on 1000 Monte Carlo samples

Source of difference			Sample size		
x	Y	Z	n_1	n_2	Power
(0) $\Delta_{12} = \Delta_{21} = 0$					
No	No	No	50	25	0.112
No	No	No	50	100	0.109
No	No	No	100	100	0.054
No	No	No	100	150	0.048
(a) $p_1 = 0.5, p_2 = 0.75$					
Yes	No	No	50	25	0.201
Yes	No	No	50	100	0.3
Yes	No	No	100	100	0.481
Yes	No	No	100	150	0.579
(b) $\mu_{11} = 0, \mu_{21} = 0.5$					
No	Yes	No	50	25	0.146
No	Yes	No	50	100	0.193
No	Yes	No	100	100	0.275
No	Yes	No	100	150	0.371
(c) $\mu_{11}^* = 0, \mu_{21}^* = 0.5$					
No	No	Yes	50	25	0.126
No	No	Yes	50	100	0.217
No	No	Yes	100	100	0.324
No	No	Yes	100	150	0.365
(d) *Differences in all three variables*					
Yes	Yes	Yes	50	25	0.287
Yes	Yes	Yes	50	100	0.483
Yes	Yes	Yes	100	100	0.733
Yes	Yes	Yes	100	150	0.849

Note: The parameters under $H : \Delta_{12} = \Delta_{21} = 0$ (i.e., under case (0)) are $p_1 = p_2 = 0.5, \mu_{11} = \mu_{22} = \mu_{11}^* = \mu_{21}^* = 0, \mu_{12} = \mu_{22} = \mu_{12}^* = \mu_{22}^* = 0.5$ with $\sigma = 1, \rho = 0.5$.

Patients were grouped according to sex (i.e., male or female), and the interest is to see whether there is a difference between these two groups. Only those subjects with waiting times to surgery exceeding 0 but not exceeding 60 h were included in the analysis. In addition, the waiting times to surgery were transformed using their natural logarithms. Normal probability plots of the transformed waiting times indicate that the assumption of normality is satisfied.

The data set is summarized with respect to the discrete variables **x** and Z in Table 2. The values of the continuous variable Y are not shown in the table but can be obtained from Fisher and Van Bell [7, p. 680]. The general mixed-data model was fit to this data set and MLEs of the parameters were calculated. These estimates are presented in Table 3 with their corresponding large-sample standard errors.

From Table 3, $\hat{\Delta}_{12}$ is found to be equal to 0.0396, and upon comparison with the 5% level critical value 14.1 obtained from the χ^2 distribution with seven degrees of

Table 2
Three-dimensional array for the appendicitis data [8]

	Males		Females		
	Perforation		Perforation		
Duration	Yes	No	Yes	No	Total
>24 h	20	26	8	14	68
$\leqslant 24$ h	5	61	5	42	113
Total	25	87	13	56	181

Note: Shown are the numbers of surgery patients classified according to population (male or female), perforation state ($s = 1$ if perforation is present and $s = 2$ otherwise), and duration ($Z = 2$ if duration exceeds 24 h and $Z = 1$ otherwise). The actual values of the time Y from diagnosis to surgery are found in [7, p. 680].

Table 3
Maximum likelihood estimates of parameters of general mixed-data model for the appendicitis data

Parameter	Male population	Female population
$\hat{p}$	0.2232 (0.039)	0.1884 (0.047)
$\hat{\mu}_1$	1.2154 (0.21)	1.1908 (0.266)
$\hat{\mu}_2$	1.5513 (0.112)	1.5972 (0.143)
$\hat{\gamma}$	0.9022 (0.174)	1.0555 (0.276)
$\hat{\beta}$	0.2448 (0.099)	0.2379 (0.144)
$\hat{\tau}$	1.4365 (0.271)	1.0512 (0.237)
	$\hat{\sigma} = 1.0535$ (0.116)	
	$\hat{\Delta}_{12} = \hat{\Delta}_{21} = 0.0396$	

Note: Shown are the maximum likelihood estimates of the general mixed-data model parameters for the male and female populations. The numbers in parentheses are the standard errors of the estimates. Also shown is the estimated generalized Mahalanobis distance between the two groups.

freedom, the test fails to reject the null hypothesis H that there is no difference due to sex. This conclusion agrees with those of Nakanishi [18].

References

[1] A.A. Afifi, R.M. Elashoff, Multivariate two sample tests with dichotomous and continuous variables. I. The location model, Ann. Math. Statist. 40 (1969) 290–298.
[2] J.A. Anderson, J.D. Pemberton, The grouped continuous model for multivariate ordered categorical variables and covariate adjustment, Biometrics 41 (1985) 875–885.
[3] A. Bar-Hen, J.J. Daudin, Generalization of the Mahalanobis distance in the mixed case, J. Multivariate Anal. 53 (1995) 332–342.
[4] E.J. Bedrick, J. Lapidus, J.F. Powell, Estimating the Mahalanobis distance from mixed continuous and discrete data, Biometrics 56 (2000) 394–401.

[5] A.R. de Leon, K.C. Carriére, General mixed-data model: extension of general location and grouped continuous models, Technical Report 02.05 Statistics Centre, Department of Mathematical and Statistical Sciences, University of Alberta, 2002. http://www.stat.ualberta.ca/stats_centre/tech.htm.
[6] F. Drasgow, Polychoric and polyserial correlations, in: S. Kotz, N.L. Johnson (Eds.), Encyclopedia of Statistical Sciences, Wiley, New York, 1986, pp. 68–74.
[7] L.D. Fisher, G. Van Bell, Biostatistics: A Methodolgy for the Health Sciences, Wiley, New York, 1993.
[8] T.D. Koepsel, T.S. Inui, V.T. Farewell, Factors affecting perforation in acute appendicitis, Surg. Gynecol. Obstet. 153 (1981) 508–510.
[9] E. Krusińska, A valuation of state of object based on weighted Mahalanobis distance, Pattern Recognition 20 (1987) 413–418.
[10] W.J. Krzanowski, Distance between populations using mixed continuous and categorical variables, Biometrika 70 (1983) 235–243.
[11] W.J. Krzanowski, On the null distribution of distance between two groups, using mixed continuous and categorical variables, J. Classification 1 (1984) 243–253.
[12] S. Kullback, Information Theory and Statistics, 2nd Edition, Dover, New York, 1968.
[13] T.W. Kurczyński, Generalized distance and discrete variables, Biometrics 26 (1970) 525–534.
[14] J. Lapidus, Multivariate statistical methods using continuous and discrete data, Ph.D. Thesis, University of New Mexico, 1998.
[15] S.-Y. Lee, W.-Y. Poon, P.M. Bentler, Simultaneous analysis of multivariate polytomous variates in several populations, Psychometrika 54 (1989) 63–73.
[16] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, New York, 1979.
[17] K. Matusita, Decision rule, based on distance, for the classification problem, Ann. Inst. Statist. Math. 16 (1956) 305–315.
[18] H. Nakanishi, Distance between populations in a mixture of categorical and continuous variables, J. Japan Statist. Soc. 26 (1996) 221–230.
[19] I. Olkin, R.F. Tate, Multivariate correlation models with mixed discrete and continuous variables, Ann. Math. Statist. 32 (1961) 448–465 (correction in 36 (1961) 343–344).
[20] W.-Y. Poon, S.-Y. Lee, Statistical analysis of continuous and polytomous variables in several populations, British J. Math. Statist. Psych. 45 (1987) 139–149.
[21] G.A.F. Seber, Multivariate Observations, Wiley, New York, 1984.