Distance Functions for Categorical and Mixed Variables

Brendan McCane * Michael Albert

Dept Computer Science, University of Otago, PO Box 56, Dunedin, New Zealand, 9015

Abstract

In this paper, we compare three different measures for computing Mahalanobis-type distances between random variables consisting of several categorical dimensions or mixed categorical and numeric dimensions - regular simplex, tensor product space, and symbolic covariance. The tensor product space and symbolic covariance distances are new contributions. We test the methods on two application domains - classification and principal components analysis. We find that the tensor product space distance is impractical with most problems. Over all, the regular simplex method is the most successful in both domains, but the symbolic covariance method has several advantages including time and space efficiency, applicability to different contexts, and theoretical neatness.

Key words: Categorical data, Mahalanobis distance.

* Corresponding Author
Email addresses: `mccane@cs.otago.ac.nz` (Brendan McCane), `malbert@cs.otago.ac.nz` (Michael Albert).

1 Introduction

In this paper, we compare three different measures for computing Mahalanobis-type distances between random variables consisting of several categorical dimensions or mixed categorical and numeric dimensions - regular simplex, tensor product space, and symbolic covariance. In each case, distances are computed via an interpretation of the categorical data in some real vector space. For carrying out practical computations, the dimension of this space is important, and the lower the dimension, the easier the computations will be. The regular simplex method is well known and involves replacing a single k level categorical variable with a $(k-1)$-dimensional numerical variable in such a way that each level of the categorical variable is mapped to a vertex of a regular simplex, and is thereby at the same distance from every other level of that variable. For a categorical variable with 3 levels, this results in an equilateral triangle in R^2. The dimension of the embedding space is thus the sum of the number of levels of all the variables, minus the number of variables. The tensor product space method is commonly used in the location model for dealing with mixed variables (Kurzanowski, 1993) although here we develop a new derivation which gives rise to a Mahalanobis-type distance in the product space. The dimension of the embedding space is the product of the number of levels of all the variables. The final method seems to be completely new and calculates an analogue of the covariance between any two categorical variables, which is used to create a Mahalanobis-type distance. In this case, strictly speaking, there is no embedding space, but all computations take place in a dimension equal to the number of variables. For comparison purposes, we also compare the results in a classification experiment with the Heterogeneous Value Difference Metric (HVDM) (Wilson and Martinez, 1997) and with Naive Bayes.

Why do we need to develop Mahalanobis-type distances for categorical and mixed variables? The chief reason is that most research on calculating distances of this sort is either heuristic (Wilson and Martinez, 1997; Gower, 1971; Cost and Salzberg, 1993; Huang, 1998), assumes the variables are independent (Kurzanowski, 1993; Bar-Hen and Daudin, 1995; Cuadras et al., 1997; Goodall, 1966; Li and Biswas, 2002), or makes use of the special case of ordinal data (Podani, 1999) by assuming an underlying distribution and a discretisation function. The Value Difference Metric (VDM) is one of the most popular categorical distances and was introduced by Stanfill and Waltz (1986) and is particular to classification problems. The metric is based on sample probabilities:

$$vdm(x_a, y_a) = \left(\sum_c P(c|x_a)^2\right)^{0.5} \sum_c |P(c|x_a) - P(c|y_a)|^2 \qquad (1)$$

where vdm is the value difference metric on attribute a, $P(c|x_a)$ is the probability of class c given that attribute a has the value x_a. The first term is a weighting term given by Stanfill and Waltz (1986) to appropriately weight the importance of each attribute. The weighting term does not appear in the description given by Wilson and Martinez (1997). The total VDM is then the sum of vdm over all attributes a.

The most influential study on mixed distances is that of Wilson and Martinez (1997). They introduce several new metrics based on the VDM: Heterogeneous Value Difference Metric (HVDM), Interpolated VDM (IVDM) and Windowed VDM (WDVM). The HVDM is very similar to the similarity metric of Gower (1971) and is used for comparison with other metrics in our study. The IVDM is an extension of the VDM that takes into account continuous attributes by discretising them to calculate sample probabilities to use in Equation 1. However, the actual probabilities used are interpolated between neighbouring bins, depending on where in the bin the continuous value actually falls. The WVDM is a more sophisticated version of IVDM where the probabilities are calculated at each point which occurs in the training set using a sliding window similar to a Parzen window Parzen (1962). The simplified VDM (without the initial weighting term) and related measures have been used in several influential works including Cost and Salzberg (1993) and Domingos (1996).

The limitations of these techniques is that they are constrained to be used in a classification context and almost all assume that attributes are independent of each other. Thus, two highly correlated attributes will contribute twice as much to the evidence as they should. We attempt to avoid these problems in the techniques presented here. Also, we are not just interested in distances between populations as in Kurczynski (1970), but also distances between individuals and distances between an individual and a population. The former is useful for clustering algorithms, and the latter for classification algorithms. The applications for such distances are predominantly in classification, clustering and dimensionality reduction. We give examples of classification and dimensionality reduction in the results.

2 Methods

2.1 *Regular Simplex Method*

The regular simplex method is the simplest of all the methods. The basic idea is to assume that any two distinct levels of a categorical variable are separated by the same distance. To achieve this, each level of an n-level variable is associated with a distinct vertex of a regular simplex in $(n-1)$ dimensional space. For simplicity, the distance between levels is assumed to be 1. For example, given a categorical variable $X \in \{A, B, C\}$, A could be mapped to $(0, 0)$, B to $(1, 0)$ and C to $(1/2, \sqrt{3}/2)$. The choice of simplex is arbitrary, and has no effect on the subsequent analysis. One possible choice is to continue the development of the sequence above. Take the existing set of vertices, and append their centroid. By the symmetry of the regular simplex, this point is at the same distance from all the preceding vertices. Now add one more coordinate, taking the value 0 on the original set of points, and set its value on the new point to be such that the distance from any, and hence all, other elements of the set is 1.

Each level of a variable is then replaced by the corresponding vertex in the simplex. For a problem with c categorical dimensions, where the k^{th} variable has n_k levels, an observation of the c-dimensional variables is replaced with a variable with $\sum_{k=1}^{c}(n_k - 1)$ numeric dimensions. A distance function can then be defined based on the covariance matrix of the replaced data points:

$$d_{rs}(x_1, x_2) = (x_1' - x_2')^T \Sigma_{rs}^{-1} (x_1' - x_2') \tag{2}$$

where x_1' is the regular simplex representation of the input vector x_1. Let $S_L = \sum_{k=1}^{c} n_k$, the sum of the levels. Σ_{rs} is of size $((S_L - c) \times (S_L - c))$ and can be naively [1] calculated in time $O(N_s S_L^2)$ where N_s is the number of samples in the data set. Σ_{rs}^{-1} can be calculated in time $O(S_L^3)$ [2]. Therefore the space complexity of this method is $O(S_L^2)$ and the time complexity is $O(N_s S_L^2 + S_L^3)$.

[1] by naive, we mean without using optimizations such as sparse matrices and sparse matrix multiplication

[2] actually in time $O(S_L^{2.376})$ using a better bound for matrix multiplication

2.2 Tensor Product Space Method

The tensor product space method is most similar in spirit to the original Mahalanobis distance derivation. The underlying idea of the Mahalanobis distance is that we wish to calculate the Euclidean distance between two n-dimensional points, p_1, p_2 where each dimension is independent of the others. Unfortunately, p_1 and p_2 cannot be measured directly, but observations q_1 and q_2, which are linear transformations of the original values, can be. That is, for some matrix A:

$$q_1 = Ap_1$$

We want to calculate the distance between p_1 and p_2 so:

$$\begin{aligned} d(p_1, p_2) &= (p_1 - p_2)^T (p_1 - p_2) && (3) \\ &= (A^{-1}q_1 - A^{-1}q_2)^T (A^{-1}q_1 - A^{-1}q_2) && (4) \\ &= (q_1 - q_2)^T {A^{-1}}^T A^{-1} (q_1 - q_2) && (5) \end{aligned}$$

The matrix ${A^{-1}}^T A^{-1}$ is just the inverse covariance matrix of the population of p's, and we're left with the classic Mahalanobis distance.

Now consider a random variable X which is defined over a space of c categorical variables where the k^{th} variable has n_k levels. For two categorical variables to be independent, the product of the marginal distributions must equal the actual distribution. That is:

$$P(X_i = a, X_j = b) = P(X_i = a)P(X_j = b).$$

The joint distribution $P(X_i, X_j)$ and the subsequent marginal distributions $P(X_i)$ can be estimated from the sample population. The joint distribution may not be independent, and to mimic the construction above we need to find a transformation from the dependent joint distribution to an independent joint distribution. The independent joint is estimated simply as the product of the marginals.

We are left with the problem of estimating a linear transformation between tensor product spaces. The initial probability tensor space is a dependent observable space, T^D:

$$T^D \cong X_1 \otimes X_2 \otimes X_3 \dots$$

where $T_i^D = P(X_1 = a, X_2 = b, X_3 = c, \dots)$. For example, with a two-dimensional random variable where the first dimension has 2 levels and the second has 3, we would get a tensor space of 6 dimensions. We want a linear transformation from T^D to T^I where T^I is the independent tensor space $T_i^I = P(X_1 = a)P(X_2 = b)P(X_3 = c) \dots$. The problem is ill-posed so there are many possible solutions. We have chosen the solution which produces a transformation as close as possible to the identity. Since both tensor product spaces are probability spaces, the transformation matrix, M, must be a column stochastic matrix, which can be defined as:

$$M_{ii} = \begin{cases} 1 & \text{if } s_i < t_i, \\ \frac{t_i}{s_i} & \text{otherwise} \end{cases}$$

$$M_{ij} = \begin{cases} \frac{t_i - M_{ii} s_i}{s_j} & \text{if } 1 - M_{jj} \geq \frac{t_i - x_{ii} s_i}{s_j}, \\ 1 - \sum_{k=1}^{j} M_{kj} & \text{otherwise} \end{cases}$$

Where s_i and t_i are the joint probabilities of the dependent and independent tensor product spaces respectively.

The matrix, M, is a transformation from a dependent space to an independent one, and as such is analogous to A^{-1} in equation 5. By analogy, we can call the matrix $\Sigma_{tp}^{-1} = M^T M$ the inverse covariance matrix of the categorical variables. A Mahalanobis-like distance function can then be defined:

$$d_{tp}(x_1, x_2) = (x_1' - x_2')^T \Sigma_{tp}^{-1} (x_1' - x_2'), \tag{6}$$

where x_1' is the tensor product space representation of x_1. Let $P_L = \prod_{k=1}^{c} n_k$, the product of the levels. Matrix M is of size $P_L \times P_L$, therefore the naive size complexity of this method is $O(P_L^2)$ and the naive time complexity is $O(P_L^3)$.

2.3 Symbolic Covariance

Consider the formula for covariance of two field-valued (generally real-valued) variables X and Y:

$$\sigma^2(X, Y) = \mathbf{E}((X - \bar{X})(Y - \bar{Y})),$$

where both **E** and an overbar indicate expectation. Now consider two categorical random variables A and B with values A_1 through A_n and B_1 through B_m respectively. For $1 \leq i \leq n$ and $1 \leq s \leq m$ let:

$$\begin{aligned}
p_{is} &\stackrel{\text{defn}}{=} \mathbf{P}(A = A_i, B = B_s) \\
a_i &\stackrel{\text{defn}}{=} \mathbf{P}(A = A_i) \\
b_s &\stackrel{\text{defn}}{=} \mathbf{P}(B = B_s).
\end{aligned}$$

Consider A_1 through A_n and B_1 through B_m as symbolic variables and define:

$$\begin{aligned}
\bar{A} &\stackrel{\text{defn}}{=} a_1 A_1 + a_2 A_2 + \cdots + a_n A_n \\
\bar{B} &\stackrel{\text{defn}}{=} b_1 B_1 + b_2 B_2 + \cdots + b_m B_m.
\end{aligned}$$

Where $\bar{A}$ and $\bar{B}$ are also symbolic expressions. Then we have:

$$A_j - \bar{A} = \sum_{i=1}^{n} a_i (A_j - A_i)$$

As A_j and A_i are categories and not values, the term $A_j - A_i$ doesn't really make sense, so we replace it with a more generic term, δ which we call the distinction between two categorical values:

$$A_j - \bar{A} \stackrel{\text{defn}}{=} \sum_{i=1}^{n} a_i \delta(A_j, A_i). \tag{7}$$

We define δ to have the following properties:

$$\delta(A_i, A_i) \stackrel{\text{defn}}{=} 0 \quad (8)$$

$$\delta(A_i, A_j) \stackrel{\text{defn}}{=} -\delta(A_j, A_i) \quad (9)$$

The definition in 9 is required so the expression $X - \bar{X}$ can be positive or negative as is the case with numeric attributes. Without this definition, the symbolic covariance could never be 0 - even if the variables are uncorrelated. Also, without this definition, the symbolic covariance would not be invariant to a reordering of categories (see Property 2 below). We note that a side effect of this definition is that the symbolic covariance could be either positive or negative as in the case of numeric covariance. However, the positivity or negativity is somewhat arbitrary due to the ordering of the categories - if negative covariances are not needed, the absolute value of the symbolic covariance can be taken.

The underlying motivation is that we view the expression $X - \bar{X}$ as representing the "difference" in the sense of "being different from" an observation of the random variable X and its mean, rather than the same "difference" in the sense of "a value computed by the rules of arithmetic". While, for real valued variables, it makes perfect sense to collapse these two meanings, this collapse is not at all self-evident, nor necessarily desirable for categorical ones.

We now propose the symbolic covariance:

$$
\begin{aligned}
\sigma_s^2(A,B) &\overset{\text{defn}}{=} \mathbf{E}((A-\bar{A})(B-\bar{B})) &(10)\\
&= \sum_{i=1}^{n}\sum_{s=1}^{m} p_{is} \sum_{j=1}^{n}\sum_{t=1}^{m} a_j b_t \delta(A_i,A_j)\delta(B_s,B_t) &(11)\\
&= \sum_{i<j}\sum_{s<t}(p_{is}a_j b_t - p_{js}a_i b_t - p_{it}a_j b_s + p_{jt}a_i b_s)\delta(A_i,A_j)\delta(B_s,B_t). &(12)
\end{aligned}
$$

This remains a symbolic expression. We can realise an actual value for σ_s^2 by choosing appropriate values of $\delta(A_i, A_j)$. In the absence of other information, choosing $\delta(A_i, A_j) = 1$ for $i < j$ and -1 for $i > j$ is a reasonable assumption. For ordinal variables, we might choose a distinction based on the ordering.

Okada (2000) apparently (he did not provide details in his paper) almost discovered σ_s^2 but it seems he failed to realise the necessity of setting $\delta(A_i, A_j) = -\delta(A_j, A_i)$.

Aside from the pragmatic possibility of using this symbolic covariance as an ingredient in a Mahalanobis-type distance, it has a number of attractive properties.

Property 1 (Independence) *If A and B are independent, then $p_{ij} = a_i b_j$ and it follows immediately that $\sigma_s^2(A, B) = 0$.*

Property 2 (Renaming) *The quantity $\sigma_s^2(A, B)$ is invariant under a renaming of the categories.*

This claim is easily verified by direct computation in the case where A_i and A_{i+1} are exchanged. The sign changes in δ expressions are exactly balanced by the reordering of the terms in their multipliers.

Property 3 ($\sigma_s^2(A, A)$) *If A is a categorical variable with n levels, then the quantity $\sigma_s^2(A, A)$ is maximised when each level is equally likely.*

We also note that the symbolic covariance has some similarities to the χ^2 statistic. While χ^2 may be used as a proxy for covariance, its purpose is very different. It essentially asks what is the probability that two (or more) variables are independent, and this is not the same as asking to what extent two variables are independent.

The symbolic covariance is technically only defined between two categorical variables, however since we can use equation 7 to transform a categorical variable to a (mean-shifted) real number, we can use the standard definition of covariance for calculating the covariance between a categorical variable and a numeric one. We show the results of a mean-shift in the examples below. We note what looks like a paradox in calculating $A_i - \bar{A}$, in particular:

$$(A_i - \bar{A}) - (A_j - \bar{A}) \neq \delta(A_i, A_j)$$

In effect, the function $A_i - \bar{A}$ as defined in equation 7 is more like a projection operator than a one-dimensional difference operator.

A Mahalanobis-like distance function can be defined using the symbolic covariance matrix:

$$d_{sc}(x_1, x_2) = (\Delta(x_1 - x_2))^T \Sigma_{sc}^{-1} \Delta(x_1 - x_2) \tag{13}$$

where $\Delta(x_1 - x_2)$ implies applying δ to each dimension in the vector. Σ_{sc} is of size $c \times c$ and since calculating the covariance between two variables takes $n_i n_j$, calculating the covariance matrix takes $O(S_L^2)$. Therefore the size complexity of the method is $O(c^2)$ and the time complexity is $O(S_L^2 + c^3)$.

2.3.1 Some Simple Examples

Let us consider some simple examples to show the utility of the method. Table 1 shows four samples from a population each with four binary attributes or variables. Looking at the variables we would expect that A and B are perfectly correlated (either positively or negatively), A and C are uncorrelated, and A and D are partially correlated.

Sample	A	$A_i - \bar{A}$	B	$B_i - \bar{B}$	C	$C_i - \bar{C}$	D	$D_i - \bar{D}$	σ_s^2
1	A_1	0.5	B_2	-0.5	C_1	0.5	D_1	0.25	$\sigma_s(A, A) = 1.0$
2	A_2	-0.5	B_1	0.5	C_2	-0.5	D_2	-0.75	$\sigma_s(A, B) = -1.0$
3	A_1	0.5	B_2	-0.5	C_2	-0.5	D_1	0.25	$\sigma_s(A, C) = 0$
4	A_2	-0.5	B_1	0.5	C_1	0.5	D_1	0.25	$\sigma_s(A, D) = 0.5$

Table 1

Some example categorical variables

To calculate the symbolic covariance for variable A, we can first shift to the mean:

$$
\begin{aligned}
A_1 - \bar{A} &= A_1 - (a_1 A_1 + a_2 A_2) \\
&= A_1 - (0.5A_1 + 0.5A_2) \\
&= 0.5A_1 - 0.5A_1 + 0.5A_1 - 0.5A_2 \\
&= 0.5\delta(A_1, A_1) + 0.5\delta(A_1, A_2) \\
&= 0.5 \\
A_2 - \bar{A} &= A_2 - (0.5A_1 + 0.5A_2) \\
&= 0.5A_2 - 0.5A_1 + 0.5A_2 - 0.5A_2) \\
&= 0.5\delta(A_2, A_1) + 0.5\delta(A_2, A_2) \\
&= -0.5
\end{aligned}
$$

The results of similar calculations are shown in every other column of Table 1. Calculating the symbolic covariance is then straightforward:

$$
\begin{aligned}
\sigma_s(A, B) &= E((A - \bar{A})(B - \bar{B})) \\
&= (0.5)(-0.5) + (-0.5)(0.5) + (0.5)(-0.5) + (-0.5)(0.5) \\
&= -1.0
\end{aligned}
$$

The results of similar calculations are shown in the last column of Table 1. Note that the results are exactly what we would intuitively expect. Any classification method that uses the notion of a distance function can then be used.

2.4 Heterogeneous Value Distance Metric (HVDM)

Although only applicable to classification problems the HVDM (Wilson and Martinez, 1997) has been influential in the literature and is included here as a comparison with the other methods. The HVDM is given by:

$$HVDM(x, y) = \sqrt{\sum_{a=1}^{m} d^2(x_a, y_a)}$$

where m is the number of attributes, and:

$$d(x_a, y_a) = \begin{cases} 1 & \text{if } x_a \text{ or } y_a \text{ is unknown,} \\ normalized\_vdm(x_a, y_a) & \text{if } a \text{ is categorical,} \\ \textit{normalized\_diff}(x_a, y_a) & \text{if } a \text{ is numeric.} \end{cases}$$

As given by Wilson and Martinez (1997), $normalized\_vdm$ is:

$$normalized\_vdm(x_a, y_a) = \sqrt{\sum_{c=1}^{C} |P(c|x_a) - P(c|y_a)|^2}$$

and $normalized\_diff$ is:

$$normalized\_diff(x_a, y_a) = \frac{|x_a - y_a|}{4\sigma_a},$$

where σ_a is the standard deviation of numeric attribute a.

3 Results

We compare the regular simplex and symbolic covariance methods on two example applications - classification and principal components. We do not include the tensor product space in the results as the technique is impractical for most problems - the dimensionality is huge for most practical problems with the result that the data set cannot be represented in machine RAM in most cases.

3.1 Classification Results

In our first experiment, we compare the performance of the two methods on example datasets with only categorical variables or with mixed categorical and numeric variables. We use the Naive Bayes classifier as the default comparison method, and the nearest neighbour classifier using HVDM as a widely used method from the literature. Since Naive Bayes has strong assumptions regarding the independence of the variables, we would expect that removing the correlation between the variables would result in improvements in many cases. We have used the discriminant analysis family of methods to test our distance calculations since these methods naturally make use of Mahalanobis like distances. The classification methods we use are:

NB: Naive Bayes used on the raw variables. If numeric variables are involved, then a normal distribution is used to model them.

LDASC: Linear Discriminant Analysis using symbolic covariance.

LDARS: Linear Discriminant Analysis using a regular simplex.

QDASC: Quadratic Discriminant Analysis using symbolic covariance.

QDARS: Quadratic Discriminant Analysis using a regular simplex.

LDASC: Regularised Discriminant Analysis using symbolic covariance.

RDARS: Regularised Discriminant Analysis using a regular simplex.

HVDMNN: Nearest Neighbour algorithm using the HVDM.

Note that for the RDA methods, the regularised parameters are not estimated using cross-validation as suggested by Friedman (1989). The parameter γ is arbitrarily set to 0.1 and the parameter λ is estimated as:

$$\lambda = 1.0 - s_i/s,$$

where s_i is the sum of the singular values in the class covariance matrix Σ_i and s is the sum of the singular values in the pooled covariance matrix Σ. LDA, QDA and RDA all use covariance matrices to form decision boundaries, so we simply use the appropriate covariance matrix within each of these methods. See McLachlan (2004) for a full description of LDA, QDA and RDA.

The results for each method on a subset (each problem has at least some categorical variables) of the UCI machine learning database (D.J. Newman and Merz, 1998) are shown in Table 2. It is not our intention to develop the best classifier for these problems although we note that many of the results are equal to the best in the literature (e.g. mushrooms), but rather to show that these methods can be useful and are therefore worth considering when dealing with categorical data. It is worth noting that sometimes QDA fails miserably when an accurate estimate of the class specific covariance matrix cannot be obtained (for example, when only one or two samples of that class exist in the data - this happens for audiology and autos). Naive Bayes performs quite well in most cases, even when the assumptions of independence are violated. On balance, the regular simplex method outperforms the method of symbolic covariance performing best in 8 problems compared with 4 for symbolic covariance. The symbolic covariance method can be likened to an a priori (sub-optimal) projection, whereas the regular simplex method can result in a more optimal data-driven projection. The symbolic covariance method performs comparably to Naive Bayes and HVDM using nearest neighbour.

	NB	LDASC	LDARS	QDASC	QDARS	RDASC	RDARS	HVDMNN
hayesRoth	78 ± 6.1	48 ± 13	75 ± 7.9	48 ± 17	43 ± 9.7	52 ± 14	$\mathbf{84 \pm 9.2}$	68 ± 9.5
lungCancer	43 ± 22	$\mathbf{63 \pm 19}$	53 ± 32	53 ± 28	40 ± 34	53 ± 17	43 ± 27	43 ± 22
promoter	$\mathbf{92 \pm 9.2}$	69 ± 16	87 ± 6.7	67 ± 17	56 ± 8.4	89 ± 8.8	82 ± 17	87 ± 9.5
monks1	75 ± 9.6	67 ± 10	73 ± 6.6	$\mathbf{85 \pm 9.5}$	82 ± 7.3	78 ± 16	83 ± 9.6	80 ± 16
monks2	59 ± 11	51 ± 16	46 ± 6.7	70 ± 9.2	34 ± 11	$\mathbf{72 \pm 7.9}$	65 ± 15	51 ± 18
monks3	$\mathbf{98 \pm 4}$	84 ± 11	92 ± 6.1	82 ± 6.6	92 ± 9.2	89 ± 6.9	93 ± 6.6	86 ± 10
tictactoe	71 ± 3.6	71 ± 4.7	$\mathbf{99 \pm 1.4}$	77 ± 4.6	89 ± 2.6	75 ± 3.9	77 ± 6.4	77 ± 5
votes	93 ± 4.1	86 ± 5.4	$\mathbf{96 \pm 3.1}$	95 ± 2.2	93 ± 4.2	93 ± 4.5	92 ± 3.3	93 ± 3.4
mushroom	94 ± 0.6	97 ± 0.7	$\mathbf{100 \pm 0}$	$\mathbf{100 \pm 0}$	99 ± 0.5	99 ± 0.3	98 ± 0.3	$\mathbf{100 \pm 0}^\dagger$
audiology	73 ± 11	70 ± 10	$\mathbf{81 \pm 10}$	0 ± 0	0 ± 0	50 ± 16	56 ± 15	42 ± 37
anneal	46 ± 6.3	92 ± 3	87 ± 3.8	15 ± 6.7	28 ± 6.2	90 ± 2.7	18 ± 5.2	$\mathbf{97 \pm 1.8}$
credit	78 ± 5.2	60 ± 3.9	76 ± 3.7	56 ± 6.7	49 ± 4.9	55 ± 6.6	45 ± 3.1	$\mathbf{82 \pm 5.3}$
heart	81 ± 9.9	83 ± 7.4	$\mathbf{84 \pm 6.6}$	80 ± 6.6	77 ± 9.2	82 ± 7.2	70 ± 8.1	79 ± 9.7
allbp	$\mathbf{96 \pm 1.2}$	89 ± 2.3	94 ± 1.7	89 ± 1.2	0.2 ± 0.3	90 ± 2	95 ± 0.8	$\mathbf{96 \pm 1.1}$
allhypo	$\mathbf{95 \pm 0.9}$	87 ± 2.1	$\mathbf{95 \pm 1.1}$	0.04 ± 0.1	0.4 ± 0.8	89 ± 1.3	89 ± 2.6	93 ± 1.9
adult	$\mathbf{55 \pm 0.7}$	33 ± 3.1	38 ± 0.7	13 ± 0.8	19 ± 1.1	35 ± 7.2	38 ± 2.4	NA‡
autos	72 ± 5.8	6.5 ± 3.4	$\mathbf{78 \pm 12}$	0 ± 0	32 ± 11	40 ± 19	32 ± 12	33 ± 17
postop	$\mathbf{67 \pm 15}$	48 ± 15	41 ± 15	1.1 ± 3.5	1.1 ± 3.5	36 ± 15	34 ± 16	51 ± 21

Table 2

Classification results on various problems from the UCI Machine Learning database. Mean and standard deviations are shown for randomised 10-fold cross-validation. The best mean value in each row is bolded. Above the line are problems with only categorical variables, below the line are problems with mixed variables. † Only 3 rounds of cross-validation were used due to large run times. ‡ No results available due to large run times.

3.2 Principal Components

For our principal components example, we use as an example multiple choice exam results from our first year programming course. We have used 3 data sets - one with 24 questions, and two with 20 questions. All answers are in the range “A” to “D” or “A” to “E”. The three data sets have 151, 157 and 200 sample points respectively. Because of the size of the resultant tensor product space, that method could not be applied to this problem. We also have the exam mark for each student and what we hope to find is that the principal component of the data is highly correlated with the mark - this should be a validation of the method. The results are shown in Figures 1 and 2. We can see from both figures that the principal component is highly correlated for both methods, thus verifying the usefulness of the techniques. However, the regular simplex method is more highly correlated and seems to be the preferable method of the two - although it comes at a cost of higher dimensionality of the problem (having 4-5 times more dimensions than the symbolic covariance method).

4 Conclusion

We have investigated the problem of distance calculations for categorical and mixed variables, and have introduced two new Mahalanobis type distances for these types of variables - the symbolic covariance method and the tensor product space method. The tensor product space method is theoretically pleasing but completely impractical for most problems. The symbolic covariance method is also theoretically pleasing, sharing many properties with the standard numeric covariance and thus leading to a natural Mahalanobis distance. It is also efficient in both time and space, can

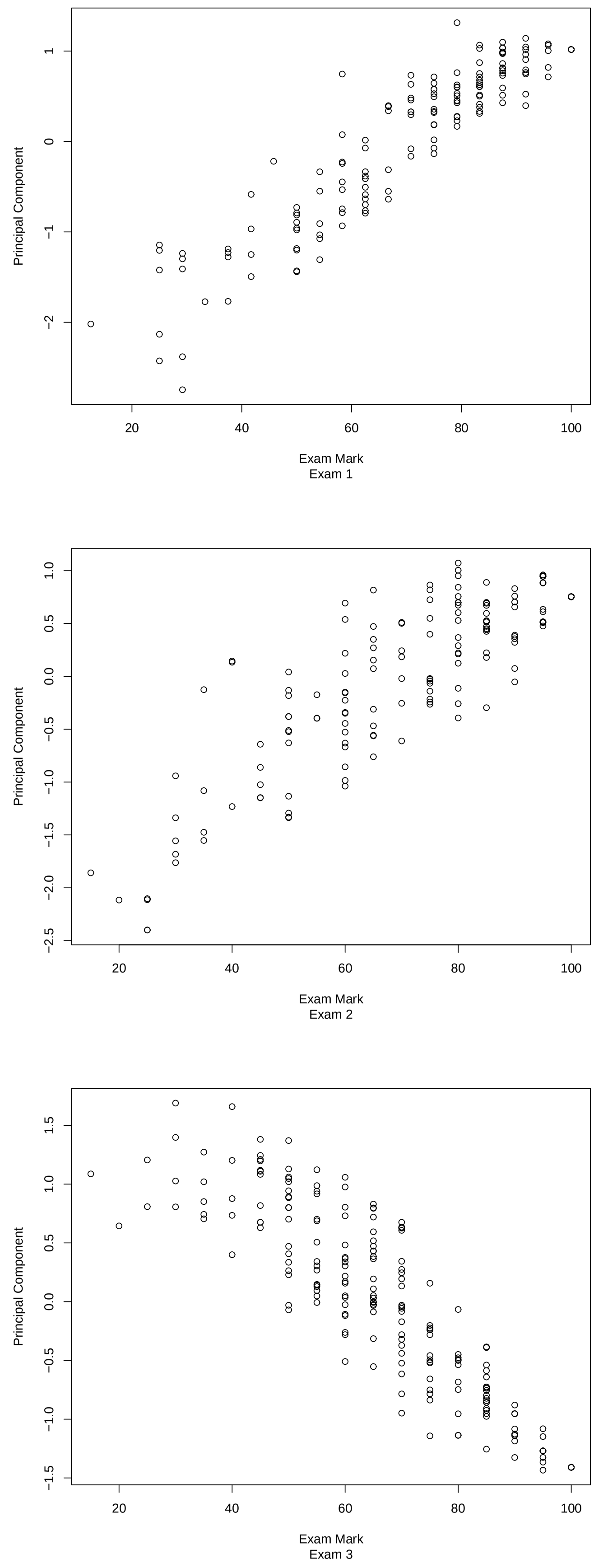

Fig. 1. Exam mark versus Principal Component for symbolic covariance.

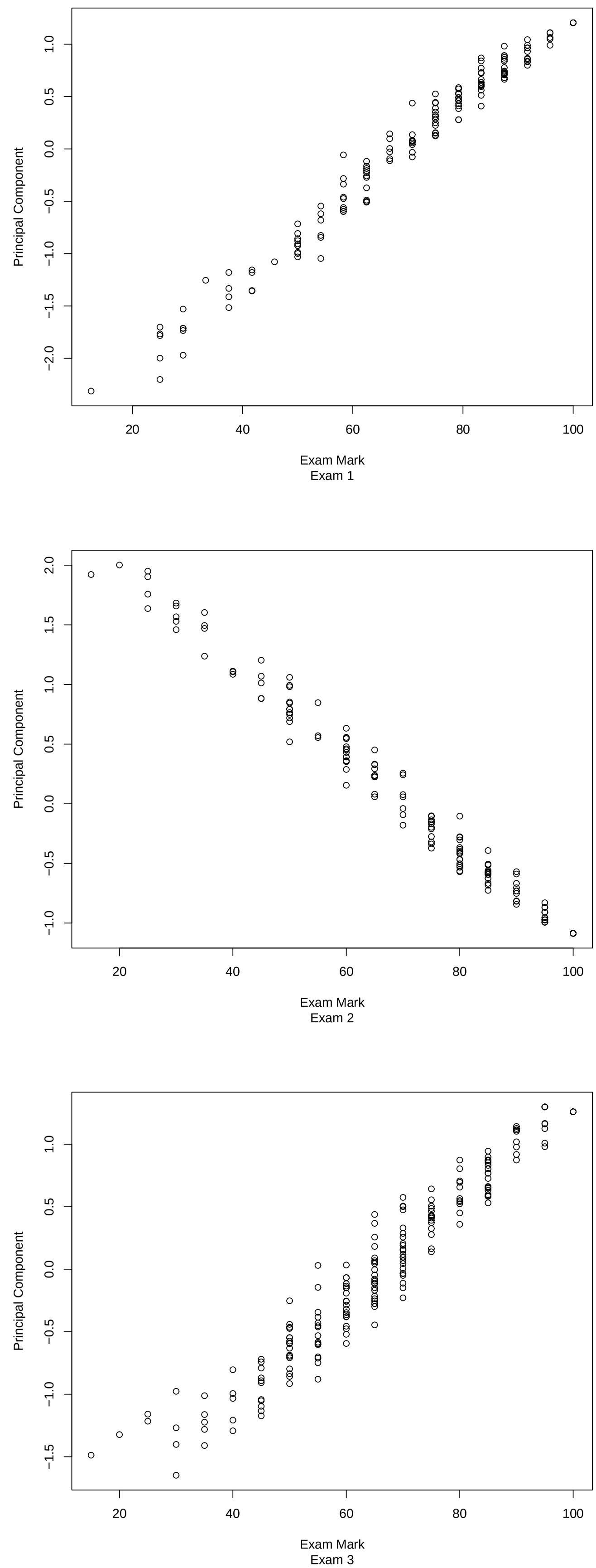

Fig. 2. Exam mark versus Principal Component for regular simplex.

be applied in many problem settings (classification, clustering and dimensionality reduction) which is not true of all methods (e.g. HVDM), and has the potential to be applied in other statistical contexts (e.g. measures of variability, statistical tests etc). Although the regular simplex method outperformed the symbolic covariance method in terms of accuracy, the performance improvement was only moderate at the cost of significantly higher dimensionality. On balance, we believe the symbolic covariance to be a useful addition to the literature on heterogeneous distance functions.

References

Bar-Hen, A., Daudin, J.-J., 1995. Generalization of the Mahalanobis distance in the mixed case. Journal of Multivariate Analysis 53, 332–342.

Cost, R. S., Salzberg, S., 1993. A weighted nearest neighbor algorithm for learning with symbolic features. Machine Learning 10, 57–78.

Cuadras, C., Forging, J., Oliva, F., 1997. The proximity of an individual to a population with applications in discriminant analysis. Journal of Classification 14 (1), 117–136.

D.J. Newman, S. Hettich, C. B., Merz, C., 1998. UCI repository of machine learning databases.
URL `http://www.ics.uci.edu/~mlearn/MLRepository.html`

Domingos, P., 1996. Unifying instance-based and rule-based induction. Machine Learning 24 (2), 141–168.

Friedman, J., 1989. Regularized Discriminant Analysis. Journal of the American Statistical Association 84 (405), 165–175.

Goodall, D., 1966. A New Similarity Index Based on Probability. Biometrics 22 (4), 882–907.

Gower, J. C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27, 857–874.

Huang, Z., 1998. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. 2 (3), 283–304.

Kurczynski, T., 1970. Generalized Distance and Discrete Variables. Biometrics 26 (3), 525–534.

Kurzanowski, W., 1993. The location model for mixtures of categorical and continuous variables. Journal of Classification 10, 25–49.

Li, C., Biswas, G., 2002. Unsupervised learning with mixed numeric and nominal data. IEEE Transactions on Knowledge and Data Engineering 14 (4), 673–690.

McLachlan, G., 2004. Discriminant Analysis and Statistical Pattern Recognition. Wiley.

Okada, T., 2000. A note on covariances for categorical data. In: K.S. Leung, L.-W. Chan, H. M. (Ed.), Intelligent Data Engineering and Automated Learning - IDEAL 2000: Data Mining, Financial Engineering, and Intelligent Agents19. Vol. 1983 / 2000 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 150–157.

Parzen, E., September 1962. On estimation of a probability density function and mode. The Annals of Mathematical Statistics 33 (3), 1065–1076.

Podani, J., 1999. Extending Gower's general coefficient of similarity to ordinal characters. Taxon 48 (2), 331–340.

Stanfill, C., Waltz, D. L., 1986. Toward memory-based reasoning. Commun. ACM 29 (12), 1213–1228.

Wilson, D. R., Martinez, T. R., 1997. Improved heterogeneous distance functions. J. Artif. Intell. Res. (JAIR) 6, 1–34.