Kernel PCA for Categorical Data

Hirotaka Niitsuma†  Takashi Okada†

† Kwansei Gakuin University, 2–1 Gakuuen, Sanda-Shi, Hyogo, 669–1337 Japan
E-mail: †{abz81166, okada}@ksc.kwansei.ac.jp

Abstract  Gini’s definition of variance for categorical data was naturally extended to covariance for mixed categorical and numerical data. In this research, we describe a procedure for calculating the covariance. Using this covariance, kernel PCA for categorical data is introduced.

Key words  Categorical data, kernel, PCA

1. Introduction

Covariances and correlation coefficients for numerical data express the strength of a correlation between a pair of variables. Such convenient measures have been expected for categorical data, and there have been many proposals to define the strength of a correlation [6]. However, none of these proposals has succeeded in unifying the correlation concept for numerical and categorical data. Recently, variance and sum of squares concepts for a single categorical variable were shown to give a reasonable measure of the rule strength in data mining [3]. If we can introduce a covariance definition for numerical and categorical variables, more flexible data mining schemes could be formulated. In this paper we propose a generalized and unified formulation for the covariance concept.

Principal Component Analysis (PCA) is an orthogonal basis transformation. The new basis is found by diagonalizing the covariance matrix. The directions of the first n Eigenvectors corresponding to the biggest n Eigenvalues cover as much variance as possible by n orthogonal directions. In many applications they contain the most interesting information: for instance, in data compression, where we project onto the directions with biggest variance to remain as much information as possible. Clearly, one cannot assert that linear PCA will always detect all structure in a given data set. By the use of suitable nonlinear features, one can extract more information. Introducing kernel function, such nonlinear features can extract from data. Such method is called Kernel PCA [1]. We apply the technique of Kernel PCA to categorical data. Using this technique, we give nonlinear extension of Gini’s variance and covariance.

2. Gini’s Definition of Variance and its Limitations

Gini successfully defined the variance for categorical data [2]. He first showed that the following equality holds for the variance of a numerical variable \( x_i \):

\[
V_i = \sum_a (x_{ia} - \bar{x}_i)^2 / n = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2
\]

(1)

where \( V_i \) is the variance of the \( i \)-th variable, \( x_{ia} \) is the value of \( x_i \) for the \( a \)-th instance, and \( n \) is the number of instances. Then, he gave a simple distance definition (1) for a pair of categorical values. The variance defined for categorical data was easily transformed to the expression at the right end of (3).
Table 1 A sample contingency table with high correlation.

<table>
<thead>
<tr>
<th>$x_i$</th>
<th>$u$</th>
<th>$v$</th>
<th>$w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>100</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>s</td>
<td>0</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>t</td>
<td>0</td>
<td>1</td>
<td>100</td>
</tr>
</tbody>
</table>

$x_{ia} - x_{ib} = \{ 
\begin{align*}
1 & \quad \text{if } x_{ia} \neq x_{ib} \\
0 & \quad \text{if } x_{ia} = x_{ib}
\end{align*}
\right.$ \hspace{1cm} (2)

$V_{ii} = \frac{1}{2n^2} \sum_{a} \sum_{b} (x_{ia} - x_{ib})^2 = \frac{1}{2} \left( 1 - \sum_{r} p_r(r)^2 \right)$ \hspace{1cm} (3)

Here $p_r(r)$ is the probability that the variable $x_i$ takes a value $r$. The resulting expression is the well-known Gini-index. The above definition can be extended to covariances by changing $(x_{ia} - x_{ib})^2$ to $(x_{ia} - x_{ib})(x_{ja} - x_{jb})$ \cite{4}. However, it does not give reasonable values relative to correlation coefficients. The difficulty can be seen in the contingency table example of Table 1. There are two variables, $x_i$ and $x_j$, each of which takes three values. Almost all instances appear in the diagonal positions, and hence the data should have a high $V_{ij}$. The problem arises when we consider an instance at $(t, v)$. Intuitively, this instance should decrease the strength of the correlation. However, there appears to be some positive contribution to $V_{ij}$ between this instance and that at $(r, u)$. It comes from the value difference pair, $(x_i : r/t, x_j : u/v)$, which is different from the major value difference pairs $(x_i : r/s, x_j : u/v)$, $(x_i : r/t, x_j : u/w)$ and $(x_i : s/t, x_j : v/w)$, and contradicts the definition of every edge is set to 1 to adapt the distance definition of \cite{2}. If there are $c$ kinds of values for a categorical variable, $x_i$, then each value can be matched to a vertex of the regular polyhedron in $(c - 1)$-dimensional space.

3.2 Definition of Covariance, $V_{ij}$

Our proposal for the $V_{ij}$ definition is the maximum value of $Q_{ij}(L_{ij})$ while changing $L_{ij}$, and $Q_{ij}(L_{ij})$ is defined by the subsequent formula,

$V_{ij} = \max(Q_{ij}(L_{ij})) \hspace{1cm} (4)$

$Q_{ij} = \frac{1}{2n^2} \sum_{a} \sum_{b} < \overline{x_{ia}x_{ib}}|L_{ij}|\overline{\overline{x_{ja}x_{jb}}}> \hspace{1cm} (5)$

Here, $L_{ij}$ is an orthogonal transformation applicable to the value space. The bracket notation, $<e|L_{ij}|f>$, is evaluated as the scalar product of two vectors $e$ and $L_{ij}f$ (or $L_{ij}^{-1}e$ and $f$). If the lengths of the two vectors, $e$ and $f$, are not equal, zeros are first padded to the vector of the shorter length.

In general, $L_{ij}$ may be selected from any orthogonal transformation, but we impose some restrictions in the following cases.

- When we compute the variance, $V_{ii}$, $L_{ii}$ must be the identity transformation, since two value difference vectors are in the identical space.
- A possible transformation of $L_{ij}$ is $(1)$ or $(-1)$ when the vector lengths of $e$ and $f$ are unity. However, if both $x_i$ and $x_j$ are numerical variables, we always have to use the transformation matrix, $(1)$, in order to express a negative correlation.

3.3 Assumed Properties for Bracket Notations

We assume several properties when using bracket notation, as follows. All these properties are easily understood as properties of a vector.

$<rr[L_{ij}]vw> = <rs[L_{ij}]uw> = 0 \hspace{1cm} (6)$

$<rs[L_{ij}]uw> = - <rs[L_{ij}]vs> \hspace{1cm} (7)$

$= - <sr[L_{ij}]uw> = <sr[L_{ij}]vu> \hspace{1cm} (8)$

$<rs[L_{ij}]uw> + <rs[L_{ij}]vw> = <rs[L_{ij}]vw> \hspace{1cm} (9)$
function [VertexList]=makingTetraVecs(c)
    d0=1;
    VertexList=[];
    v=[0];
    VertexList=[VertexList;v];
    v=[1];
    VertexList=[VertexList;v];
    for dd=d0+1:c
        vMean=mean(VertexList);
        sumNorm =0;
        for k=1:dd
            distance=norm(VertexList(k,:)-vMean);
            sumNorm=sumNorm+distance;
        end
        length=sumNorm/dd;
        height=sqrt(1-length^2) ;
        v=[vMean,height ];
        VertexList=[ [VertexList,zeros(dd,1)];v ];
    end

Fig. 2 Procedure for yielding the regular polyhedron in [c]-
dimensional

Fig.2 Procedure for yielding the regular polyhedron in [c]-
dimensional

<rs|Lij|rs> = 1 .................................. (10)

3.4 Regular polyhedron

In our definition, a categorical variable \( x_i \), are represented
by each vectors of a vertex of the regular polyhedron. When
there are \( c \) kinds of values for a categorical variable , the
regular polyhedron is a regular polyhedron in \((c-1)\)-dimensional
space. To hold the properties (10),(9), the regular polyhedral-
should be the following polyhedron.

- When \( c = 2 \), the regular polyhedron should be a seg-
- When \( c = 3 \), the regular polyhedron should be an
equilateral triangle.
- When \( c = 4 \), the regular polyhedron should be an
equilateral tetrahedron.

For the case \( c > 4 \), all faces of the regular polyhedron should
be equivalent equilateral triangular faces. Such regular poly-
hedron can be yielded by a procedure described in figure 2.
Let us represent vertexes of this polyhedron by

\[ V_i(c) = [v_i(1), v_i(2), v_i(3), \ldots v_i(c)] \]

where \( v_i(r) \) denotes a vector of a vertex represents a state
that \( x_i \) takes \( r \) th value.

3.5 Determining the orthogonal matrix: \( L \)

By using \( V_i(c) \), the optimization problem (4) can be des-
cribed as follows.

\[
\max_{L_{ij}} \text{trace}(A_{ij}L_{ij}^t) \\
\]

\[ L_{ij}L_{ij}^t = E \] .................................. (11)

where

\[ A_{ij} = \sum_{a} \sum_{b} (v_i(x_{ia}) - v_i(x_{ia}))(v_j(x_{ja}) - v_j(x_{ja}))^t. \]

\( t \) represents transpose. \( v_i(r) \) is a column-vector. Station-
ally points of a Lagrangian relaxation problem of (11) are
solutions of the following simultaneous equations.

\[ A_{ij}L_{ij} = (A_{ij}L_{ij})^t \]

\[ L_{ij}L_{ij}^t = E \] .................................. (12)

The simultaneous equations can be solved numerically. A
certain solution of the simultaneous equations is chosen as
the orthogonal matrix \( L \).

4. Nonlinear extension using Kernel

In this section, nonlinear extension of the covariance de-
ined in section 3., is discussed.

\( v_i(x_i) \) is regarded as mapping from categorical variable
\( x_i \) to real valued feature space \( U_i \subset \mathbb{R}^{c-1} \). In this feature
space, mutual distance between one category \( v_i(r) \) and an-
other category \( v_i(s) \) is always 1. This feature space is not
appropriate to some type categorical data. For example, if
\( x_i \) is ordered categorical variable, it is natural to assume the
following mapping
\[ \mathbf{v}_i(1) = 1, \mathbf{v}_i(2) = 2, \mathbf{v}_i(3) = 3, \ldots. \]
However, in this mapped feature space, the mutual distance is not always 1.

To define covariance based on Gini’s variance on such appropriate feature space \( \Theta \), we introduce nonlinear mapping \( \Phi_i \) from \( U_i \) to \( \Theta \). In section 3., relation between feature space \( U_i \) and \( U_j \) is defined by the orthogonal matrix \( L_{ij} \). We define same relation on nonlinearly mapped feature space \( \Theta \) as follows
\[ \Phi_i(\mathbf{v}_i(u)) = \sum_{a} L_{ij,a} \Phi_j(\mathbf{v}_j^a)(u), \quad (13) \]
where, \( L_{ij,a} \) a \((r,a)\) element of matrix \( L_{ij} \). Using \( \Phi_i \) and \( \Phi_j \), we can define covariance \( V_{ij} \) as follows.
\[ V_{ij} = \frac{1}{2n^2} \sum_{a} \sum_{b} \left( \Phi_i(\mathbf{v}_i(x_{ia})) - \Phi_i(\mathbf{v}_i(x_{ib})) \right) \left( \Phi_j(\mathbf{v}_j(x_{ja})) - \Phi_j(\mathbf{v}_j(x_{jb})) \right) \quad (14) \]
Then we get similar formulation.
\[ V_{ij} = \text{trace}(A_{ij}L_{ij}^t), \quad (15) \]
where
\[ A_{ij} = \frac{1}{n^2}([\text{K}_{ij}]_{r,s} - \text{1}^t[N_{ij}]_{r,s}) \]
\[ [\text{K}_{ij}]_{r,s} = \Phi_i(\mathbf{v}_i(r)) \Phi_j(\mathbf{v}_j(s)) = K_j(\mathbf{v}_j(r), \mathbf{v}_j(s)) \]
\[ [N_{ij}]_{r,s} = \sum_{a} \delta_{x_{ia}r} \delta_{x_{ja}s} \]
\[ \text{1} = (1, 1, 1, ... 1)^t \]
\( K_j(x, y) \) is kernel function. From the discussion same as section 3., the orthogonal matrix \( L_{ij} \) determined by the following optimization problem.
\[ \max_{L_{ij}} \text{trace}(A_{ij}L_{ij}^t) \]
\[ L_{ij}L_{ij}^t = \mathbf{E} \quad (16) \]

5. Kernel Principal Component Analysis (Kernel PCA)

Using definitions of covariance in section 3., 4., we can define the covariance matrix of a categorical data.
\[ C = [V_{ij}] = \begin{pmatrix} V_{11} & V_{12} & V_{13} & \ldots \\ V_{21} & V_{22} & \ldots \\ \vdots & \vdots & \ddots \end{pmatrix} \quad (17) \]
By diagonalizing the covariance matrix, PCA and Kernel PCA for a categorical data is achieved. In this section, experiments of PCA and Kernel PCA for some sample categorical data are discussed.

5.1 Samples of Covariance Matrices

There is no way to prove the proposed covariance definition. Covariance matrices are derived for typical contingency tables to facilitate the understanding of our proposal.

Our first example is the following 2 x 3 contingency table shown

\[
\begin{array}{c|ccc}
 & u & v & w \\
\hline
x_i & r & n/6 & 0 & n/6 \\
s & n/6 & n/3 & 0 & n/6 \\
\end{array}
\]
For this contingency table, PCA gives covariance matrix \( C \) is
\[ C = [V_{ij}] = \begin{pmatrix} 0.22 & 0.096 \\ 0.096 & 0.33 \end{pmatrix} \quad (18) \]
It’s eigenvalues are
\[ \lambda = (0.39, 0.17) \]
\[ \lambda_1/\lambda_2 = 2.3 \]
The correlation coefficient is
\[ R_{ij} = 0.35 \]
Kernel PCA is introduced to realize PCA on appropriate feature space. If in the above data, differences among each category are small, then, such feature can be introduced by using Gaussian kernel with large variance : 
\[ K(x, y) = \exp(-(x - y)^2/10) \]
Results of Kernel PCA using this kernel function are as follows
\[ C = [V_{ij}] = \begin{pmatrix} 0.0044 & 0.0017 \\ 0.0017 & 0.0066 \end{pmatrix} \quad (19) \]
\[ \lambda = (0.0075, 0.0035) \]
\[ \lambda_1/\lambda_2 = 2.1 \]
\[ R_{ij} = 0.31 \]
The correlation coefficient and the ratio of maximum and minimum of eigenvalues are smaller than results of normal PCA. These decreases seem to be reasonable results, because the Kernel function decreases differences among each category.

Next example is the following 2 x 3 contingency table shown

\[
\begin{array}{c|ccc}
 & u & v & w \\
\hline
x_i & r & n/3 & n/3 & 0 \\
s & 0 & 0 & 0 & n/3 \\
\end{array}
\]
In this contingency table, $x_i$ and $x_j$ have high correlation. For this contingency table, we give covariance matrix $C$ and it’s eigenvalues by using PCA.

$$C = [V_{ij}] = \begin{pmatrix} 0.22 & 0.60 \\ 0.60 & 0.33 \end{pmatrix} \quad (20)$$

The eigenvalues are:

$$\lambda = (0.48, 0.078)$$

$$\lambda_1/\lambda_2 = 6.2$$

$$R_{ij} = 0.71$$

Results of Kernel PCA using the kernel function are as follows:

$$C = [V_{ij}] = \begin{pmatrix} 0.0044 & 0.0038 \\ 0.0038 & 0.0066 \end{pmatrix} \quad (21)$$

$$\lambda = (0.0095, 0.00015)$$

$$\lambda_1/\lambda_2 = 6.3$$

$$R_{ij} = 0.71$$

The correlation coefficient and the ratio of maximum and minimum of eigenvalues are almost same. These results also seem reasonable. Because most data are on principal component, thus decreases differences among each category do not affect to the correlation coefficient and the ratio of eigenvalues.

5.2 Postoperative Patient Data

Our method can give the covariance of mixed categorical and numerical data. To execute the experiment for mixed categorical and numerical data, we select Postoperative Patient Data from the UCI repository. This data consists of 8 attributes, one numeric with missing values, objective variable consists of 3 classes, 90 instances are recorded. By using our method, covariance matrix $C$ of this data and it’s eigenvalues are calculated. Figure 3 shows eigenvalues of this data. Eigenvalues indexed 1 and 2 are much smaller than other values. From this result, we can say Postoperative Patient Data can be explained by fewer variables. Further interpretation of this result is the future work.

6. Conclusion

We discried a definition for the variance-covariance matrix that is equally applicable to numerical, categorical and mixed data. And it’s nonlinear extension is discussed. Calculations on sample contingency tables yielded reasonable results. When applied to numerical data, the proposed scheme reduces to the conventional variance-covariance concept. When applied to categorical data, it covers Gini’s variance concept. The previous work did not give an explicit algorithm to compute the variance-covariance matrix. This current work gives the explicit algorithm.


