社団法人　情報処理学会　研究報告
IPSJ SIG Technical Report

2003－ＩＣＳ－133（16）
2003／9／15

Kernel PCA for Categorical Data

HirotakaNiitsuma† TakashiOkada†

† KWANSEI GAKUIN UNIVERSITY, 2-1 GAKUEN, 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

カテゴリーデータに対するカーネル主成分分析

新妻 弘崇† and 岡田 孝†

† 関西学院大学　〒669-1337 兵庫県三田市学園 2-1
E-mail: †{abz81166,okada}@ksc.kwansei.ac.jp

あらまし　Giniによって導入された分散の定義を拡張し, 共分散をもとめる手法について議論する. この拡張によって導入される共分散は, 数値とカテゴリーデータの複合したデータに対しても適用できる. この共分散を使って主成分分析を行う. また主成分分析の非線形な拡張であるカーネル主成分分析についても議論する.

キーワード　カテゴリーデータ, カーネル, 主成分分析

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 Eigenvector corresponding to the biggest n Eigenvalues cover as much variance as possible by n orthogonal directions. In many applications they contain the most interestion information: for instance , in data compression, where we project onto the directions with biggest variance to remain as much information as possible. Clearly, one cannnot 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. Intorducing 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.

$$V_{ii} = \sum_a (x_{ia} - \bar{x_i})^2/n = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 \quad (1)$$

where V_{ii} 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).

表 1 A sample contingency table with high correlation.
Table 1 A sample contingency table with high correlation.

x_i \ x_j	u	v	w
r	100	0	0
s	0	100	0
t	0	1	100

$$x_{ia} - x_{ib} = \begin{cases} 1 & if \quad x_{ia} \neq x_{ib} \\ 0 & if \quad x_{ia} = x_{ib} \end{cases} \tag{2}$$

$$V_{ii} = \frac{1}{2n^2} \sum_a \sum_b (x_{ia} - x_{ib})^2 = \frac{1}{2}(1 - \sum_r p_i(r)^2) \tag{3}$$

Here $p_i(r)$ 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 $(x_{ia}-x_{ib})(x_{ja}-x_{jb})$ [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). This contradiction comes from (2) in that it does not discriminate between these four types of value difference pairs.

3. Generalized Covariance

We proposed a scheme to generalize the definition of a covariance for categorical data.[5]. It employs Gini' s variance definition (3) as the starting point, and introduces two additional concepts. The first is to represent the value difference as a vector in value space. The other is to regard the covariance as the extent of maximum overlap between vectors in two value spaces.

3.1 Vector Expression of a Value Difference

We employ a vector expression, $\overrightarrow{x_{ia}x_{ib}}$, instead of the distance, $x_{ia} - x_{ib}$, in the variance definition. When x_i is a numerical variable, the expression is a vector in one-dimensional space. The absolute value and sign of $(x_{ib} - x_{ia})$ give its length and direction, respectively. Now let us think of a categorical variable, x_i, that can take three values, (r s t). We can position these values at the three vertices of an equilateral triangle as shown in Figure 1. Then, a value difference is a vector in two-dimensional space. The length

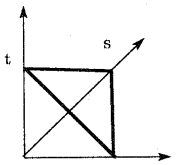


図 1 Value space.
Fig. 1 Value space.

of every edge is set to 1 to adapt the distance definition of (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,

$$V_{ij} = \max(Q_{ij}(L_{ij})) \tag{4}$$

$$Q_{ij} = \frac{1}{2n^2} \sum_a \sum_b < \overrightarrow{x_{ia}x_{ib}} | L_{ij} | \overrightarrow{x_{ja}x_{jb}} > \tag{5}$$

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 Lf (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.

$$< rr|L_{ij}|uv > = < rs|L_{ij}|uu > = 0 \tag{6}$$

$$< rs|L_{ij}|uv > = - < rs|L_{ij}|vs >$$
$$= - < sr|L_{ij}|uv > = < sr|L_{ij}|vu > \tag{7}$$

$$< rs|L_{ij}|uv > + < rs|L_{ij}|vw >$$
$$= < rs|L_{ij}|uw > \tag{8}$$

$$< rs|L_{ij}|uv > + < st|L_{ij}|uv >$$
$$= < rt|L_{ij}|uv > \tag{9}$$

```
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
```

図 2 Procedure for yielding the regular polyhedron in (c)-dimensional

Fig. 2 Procedure for yielding the regular polyhedron in (c)-dimensional

$$< rs|L_{ii}|rs >= 1 \tag{10}$$

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 polyhedron should be the following polyhedron.

- When $c = 2$, the regular polyhedron should be a segment.
- 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 polyhedron can be yielded by a procedure described in figure 2. Let us represent vertexes of this polyhedron by

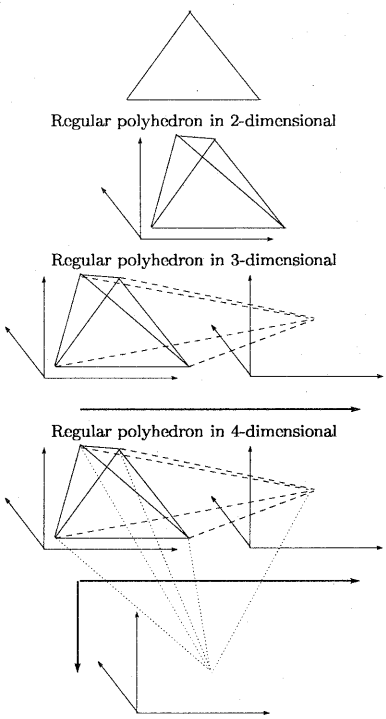
$$V_i(c) = [\mathbf{v}_i(1), \mathbf{v}_i(2), \mathbf{v}_i(3), ...\mathbf{v}_i(c)],$$

where $\mathbf{v}_i(r)$ 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 described as follows.

$$\max_{L_{ij}} \quad trace(A_{ij} {L_{ij}}^t)$$



$$L_{ij} {L_{ij}}^t = \mathbf{E} \tag{11}$$

where

$$A_{ij} = \sum_a \sum_b (\mathbf{v}_i(x_{ia}) - \mathbf{v}_i(x_{ib}))(\mathbf{v}_j(x_{ja}) - \mathbf{v}_j(x_{jb}))^t.$$

t represents transpose. $\mathbf{v}_i(r)$ is a column-vector. Stationaly points of a Lagrangian relaxation problem of (11) are solutions of the following simultaneous equations.

$$A_{ij} L_{ij}^t = (A_{ij} L_{ij}^t)^t$$
$$L_{ij} L_{ij}^t = \mathbf{E} \tag{12}$$

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 defined in section 3., is discussed.

$\mathbf{v}_i(x_i)$ is regarded as mapping from categorical variable x_i to real valued feature space $U_i \subset R^{c-1}$. In this feature space, mutual distance between one category $\mathbf{v}_i(r)$ and another category $\mathbf{v}_i(s)$ 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

$$\mathbf{v}_i(1) = 1, \mathbf{v}_i(2) = 2, \mathbf{v}_i(3) = 3, ...$$

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 Θ, we introduce nonlinear mapping Φ_i from U_i to Θ. 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 Θ as follows

$$\Phi_i(\mathbf{v}_i(r)) = \sum_u L_{ij,ru}\Phi_j(\mathbf{v}_j(u)), \tag{13}$$

where, $L_{ij,ru}$ is a (r,u) element of matrix L_{ij}. Using Φ_i and Φ_j, we can define covariance V_{ij} as follows.

$$V_{ij} = \frac{1}{2n^2}\sum_a\sum_b (\Phi_i(\mathbf{v}_i(x_{ia})) - \Phi_i(\mathbf{v}_i(x_{ib})))(\Phi_j(\mathbf{v}_j(x_{ja})) - \Phi_j(\mathbf{v}_j(x_{jb}))) \tag{14}$$

Then we get similar formulation.

$$V_{ij} = trace(A_{ij}L_{ij}^t), \tag{15}$$

where

$$A_{ij} = \frac{1}{n^2}(N_{ij}K_j^t n - N_{ij}\mathbf{1}\mathbf{1}^t N_{ij}K_j^t)$$

$$[K_j]_{r,s} = \Phi_j(\mathbf{v}_j(r))\Phi_j(\mathbf{v}_j(s)) = K_j(\mathbf{v}_j(r), \mathbf{v}_j(s))$$

$$[N_{ij}]_{r,s} = \sum_a \delta_{x_{ia},r}\delta_{x_{ja},s}$$

$$\mathbf{1} = (1, 1, 1, ...1)^t$$

$K_j(x, y)$ is kernel function. From the discussion same as section 3., the orthogonal matrix L_{ij} is determined by the following optimization problem.

$$\max_{L_{ij}} \quad trace(A_{ij}{L_{ij}}^t)$$
$$L_{ij}{L_{ij}}^t = \mathbf{E} \tag{16}$$

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.

$$C = [V_{ij}] = \begin{pmatrix} V_{11} & V_{12} & V_{13} & ... \\ V_{21} & V_{22} & ... & \\ ... & & & \end{pmatrix} \tag{17}$$

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

		x_j: u	v	w
x_i	r	n/6	0	n/6
	s	n/6	n/3	n/6

For this contingency table,PCA gives covariance matrix C is and it's eigenvalues.

$$C = [V_{ij}] = \begin{pmatrix} 0.22 & 0.096 \\ 0.096 & 0.33 \end{pmatrix} \tag{18}$$

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

$$C = [V_{ij}] = \begin{pmatrix} 0.0044 & 0.0017 \\ 0.0017 & 0.0066 \end{pmatrix} \tag{19}$$

$$\lambda = (0.0075, 0.0035)$$

$$\lambda_1/\lambda_2 = 2.1$$

$$R_{ij} = 0.31$$

The correlation coefficient and the ratio of maximum and minimum of eigenvalues are smaler 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

		x_j: u	v	w
x_i	r	n/3	n/3	0
	s	0	0	n/3

In this contingency table, x_i and x_j have high correlation. For this contingency table, we gives 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 calclated. 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 discribed 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 previouse work did not give an explicit algorithm to compute the variance-covariance matrix. This current work gives the explicit algorithm.

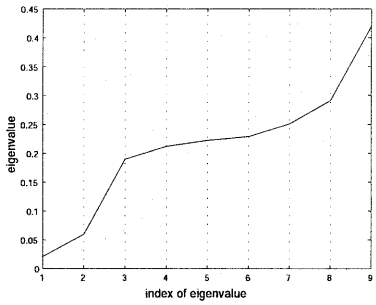


図 3 Eigenvalue-list
Fig. 3 Eigenvalue-list

文 献

[1] A. Smola B. Scholkopf and K.-R. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[2] C.W. Gini. Variability and mutability, contribution to the study of statistical distributions and relations. studi economico-giuridici della r. universita de cagliari (1912). reviewed in: Light, r.j., margolin, b.h.: An analysis of variance for categorical data. *J. American Statistical Association*, 66:534–544, 1971.

[3] T. Okada. Rule induction in cascade model based on sum of squares decomposition. In Rauch J. Zytkow, J.M., editor, *Principles of Data Mining and Knowledge Discovery (PKDD'99).LNAI*, volume 1704, pages 468–475, 1999.

[4] T. Okada. Sum of squares decomposition for categorical data. Technical report, Kwansei Gakuin Studies in Computer Science, 1999.

[5] T. Okada. A note on covariances for categorical data. In K.S. Leung, L.W. Chan, and H. Meng, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2000*, 2000.

[6] K. et al. Takeuchi. *Encyclopedia of Statistics (in Japanese)*. Toyo Keizai Shinpou Tokyo, 1990.