Compositional Data Analysis and Its Applications

Huiwen Wang
School of Economic Management, Beijing Univ. of Aeronautics and Astronautics, Beijing 100083, China

E-mail: wanghw@vip.sina.com
Outline

◆ Concept of Compositional Data
◆ Logratio Transformation for CD
◆ Predictive Modeling of CD
◆ Simple Linear Regression Model of CD
◆ Multiple Linear Regression Model of CD
I. Concept of Compositional Data

Compositional data is a very useful type of data implemented in technology, economy and social science.

For Example:

- The market share of enterprises
- The proportion of GDP in three industries
- The percentage structure within the different levels of income
- Proportion of male and female students in CNAM
AIRBUS’ MARKET IN CHINA

- 1985: AIRBUS sold the first civil airplane to China
- 1995: There were only 29 airplanes of AIRBUS served in China
- 2005: 1/6 airplanes sold in AIRBUS was destined to China (65 airplanes)
- In 2011, China Civil Airline will need about 1600 airplanes. And market share of AIRBUS would arrive at 50% in China. (Forecasting report of AIRBUS)
2. Definition of Compositional Data

Compositional Data:

Mathematically, a Pie Chart can be expressed by a vector named Compositional Data,

\[ \mathbf{X} = (x_1, x_2, \ldots, x_p)' \in \mathbb{R}^p \]

which is subject to the constraints below,

\[ \sum_{j=1}^{p} x_j = 1, \quad x_j \geq 0 \]

Hereafter we often call \( x_j, j=1,2,\ldots,p \) as “a component”
Concept of Compositional Data

Concept of composition data originally comes from work by Ferrers (1866). In 1879, Pearson discussed the complexity of its theoretical properties and indicated that in practice of compositional data analysis, “sum to unity” constraint were always been ignored consciously or unconsciously. Some traditional statistical methods designed for “unconstrained data” were often misused that led to disastrous results.

The first systematical research on compositional data was given by Aitchison (1986), which detailed studied on logistic normal distribution and logratio transformation of compositional data. Due to his special contribution in the field of compositional data, he won the Research Medal of British Royal Statistical Academy in 1988.
(1) If there exists a set of compositional data indexed by time, how to build a model hence to predict the value of each component in the future? In another words, if we have a time sequence of pie charts, how to predict pie chart in the future?

(2) How to build a simple linear regression model when both $y$ and $x$ are compositional data?

(3) If there exists one compositional data as dependent variable $y$, at meantime there exists multiple compositional data $x_1, x_2, \ldots, x_p$ as independent variables, how to construct a multiple linear regression model?
Problems

(1) Forecasting Model

An usually used method: to build models for each component separately and then to predict the ratio in the future.

Consequence: this kind of modeling method often destroy the unit-sum constraint, i.e., the sum of forecasting ratios would not be equal to 1.

(2) Regression Model

Both of the two constrains \[ \sum_{j=1}^{q} y_j = 1 \] \[ y_j \geq 0 \] would be destroyed in the predictive value of independent variable y.
II. Logratio Transformation

1. Logratio Transformation (Aitchison, 1986):

Let \( u_j = x_j / x_p \), \( j = 1, 2, \ldots, p - 1 \)

We get: \( u_1 = x_1 / x_p, \ u_2 = x_2 / x_p, \ \ldots, \ u_{p-1} = x_{p-1} / x_p, \ u_p = 1 \)

Denote \( y_j = \ln(u_j) \), then \( y_j \in (-\infty, +\infty) \)

Reason: \( X = (x_1, x_2, \ldots, x_p)' \in \mathbb{R}^p \)

0 \leq x_j \leq 1, \ j = 1, 2, \ldots, p

Let \( y_j = \ln(x_j / x_p) \), \( j = 1, 2, \ldots, p - 1 \)

And \( u_j \in [0, +\infty) \)

Delete this constrains
### Original Data

<table>
<thead>
<tr>
<th>$x_j$</th>
<th>$x_j/x_p$</th>
<th>$\ln (x_j/x_p)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>0.25</td>
<td>-1.39</td>
</tr>
<tr>
<td>0.05</td>
<td>0.63</td>
<td>-0.47</td>
</tr>
<tr>
<td>0.15</td>
<td>1.88</td>
<td>0.63</td>
</tr>
<tr>
<td>0.2</td>
<td>2.50</td>
<td>0.92</td>
</tr>
<tr>
<td>0.23</td>
<td>2.88</td>
<td>1.06</td>
</tr>
<tr>
<td>0.27</td>
<td>3.38</td>
<td>1.22</td>
</tr>
<tr>
<td>0.08</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Advantage of logratio transformation

- Firstly, it is easier to choose a function fitting for the model as the value of $y_j'$ varies in the interval $(-\infty, +\infty)$.

- Secondly it is possible to transform a nonlinear model into a linear one due to the logratio transformation.

- Thirdly, Aitchison has proved that if compositional vector $X$ follows additive logistic normal distribution, the transformed vector $Y$ will follow the normal distribution.
Inverse Transformation

Since: \[ y_j = \ln(x_j / x_p), \quad j = 1, 2 \ldots, p - 1 \]

\[ x_j = x_p e^{y_j}, \quad j = 1, 2 \ldots, p - 1 \]

Moreover: \[ x_1 + x_2 + \cdots + x_{p-1} + x_p = x_p \left( \sum_{j=1}^{p-1} e^{y_j} + 1 \right) = 1 \]

Then:

\[ x_p = \left\{ 1 / \left( 1 + \sum_{j=1}^{p-1} e^{y_j} \right) \right\} \]

\[ x_j = \left\{ e^{y_j} / \left( 1 + \sum_{j=1}^{p-1} e^{y_j} \right) \right\}, \quad j = 1, 2, \ldots, p - 1 \]
Inconvenient

(1) Short of exploitability to the modeling results

\[ y_j = \ln \left( \frac{x_j}{x_p} \right), \quad j = 1, 2, \ldots, p - 1 \]

For Example: Market Share of airplane in China

\( x_1 \) — ARIBUS, \( x_2 \) — BOENING, \( x_3 \) — OTHER

\[ y_1 = \ln \frac{x_1}{x_3}, \quad y_1 = \ln \frac{x_2}{x_3} \]

(2) 

\[ 0 < x_j^t < 1, \quad j = 1, 2, \ldots, p \]
2. Symmetrical Logratio Transformation:
(Aitchison, 1986)

\[ y_j = \log \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_i}}, \ j = 1, 2, \ldots, p \]

- The value of \( y_j \) varies in the interval \((−\infty, +\infty)\).
- Since the transformation is symmetrical to all \( x_j \), it would be easy to explain the meaning of \( y_j, j=1,2,\ldots,p \).
Inverse Transformation

Denote: \[ w_j = y_j - y_p = \ln \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_j}} - \ln \frac{x_p}{\sqrt[p]{\prod_{i=1}^{p} x_j}} = \ln \left( \frac{x_j}{x_p} \right) \]

\[ x_j = e^{w_j} x_p , \quad j = 1, 2, \ldots, p \]

Since: \[ \sum_{j=1}^{p} x_j = x_p \left( \sum_{j=1}^{p-1} e^{w_j} + 1 \right) = 1 \]

\[ x_p = \frac{1}{\sum_{j=1}^{p-1} e^{w_j} + 1} \]

\[ x_j = \frac{e^{w_j}}{\sum_{j=1}^{p-1} e^{w_j} + 1}, \quad j = 1, 2, \ldots, p - 1 \]
Drawback

Complete Collinearity within $y_j, j=1,2,\ldots p$.

\[
\sum_{j=1}^{p} y_j = \sum_{j=1}^{p} \ln \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_i}} = \ln \prod_{j=1}^{p} \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_i}} = \ln 1 = 0
\]
III. Predictive Modeling Method of Compositional Data

1. Issue

Now suppose a set of compositional data, \( X^t \), collected according to time sequence, as follows,

\[
X^t = \left\{ (x^t_1, \ldots, x^t_p)' \in \mathbb{R}^p \left| \sum_{j=1}^t x^t_j = 1, 0 \leq x^t_j \leq 1 \right. \right\}, \quad t = 1, 2, \ldots, T
\]

The objective of the study is to build a model from the given data record, so as to predict \( X^{T+l} \) at time \( T+l \),

\[
X^{T+l} = \left\{ (x^{T+l}_1, \ldots, x^{T+l}_p)' \in \mathbb{R}^p \left| \sum_{j=1}^p x^{T+l}_j = 1, 0 \leq x^{T+l}_j \leq 1 \right. \right\}
\]
The changes of three industrial structures happened in 1952, 1985 and 2000

Fig. 1  Ratios of GDP of three industries in China

Comparing the three pie charts:

- Ratio of Primary Industry in GDP decreased from 83.5% to 50.1%;
- Ratio of Secondary industry increased from 9.1% up to 26.9%;
- Ratio of Tertiary industry increased from 7.4% up to 23.0%.
Prediction to Beijing’s Employment Demand

Forecasting Objects
(Labor and Social Security Bureau of Beijing)

1. Beijing’s total employment in the next year
2. Employment of Beijing’s three industries in the next year
   - Primary ind., Secondary Ind., Tertiary ind.
3. Employment of Beijing’s five ownerships in the next year
   - State-own., collective own., Private own., Village own., Other
4. Employment amount of Beijing’s 18 districts in the next year

Divided the total amount by different structures
2. Imprecise Method

Build models for each component $x_j$ separately. For a given $j=1,2,\ldots,p$, a model is built according to the data record. Then the model is used for prediction of future data.

\[
x_j^t, \quad t = 1,2,\ldots,T \rightarrow x_j^{T+l}
\]

Consequentially, this kind of modeling method often results in failure of unit-sum constraint,

\[
\sum_{j=1}^{p} x_j^{T+l} \neq 1
\]

The reason causing the failure is that the DoF of a $p$-dimensional vector of compositional data is only $p-1$. When we build $p$ models for each component respectively, there exists a redundant of DoF such that the constraint of unit-sum could not be satisfied.
Flow chart of the modeling

compositional data $x^t_j$  

logratio transformation

non-compositional data $y^t_j$

Classical prediction method

$\hat{y}^t_j = f_j(t) + \epsilon^t_j$

predicted compositional data $x^{T+l}_{T+j}$

Inverse transformation

predicted $y^{T+l}_{T+j}$
3. Predictive Modeling (1) 
—— by use of logratio transformation

**Step 1** Perform logratio transformation defined by Aitchison on the observed compositional data as follow:

\[
y_j^t = \ln(x_j^t / x_p^t) \quad j = 1, 2, \cdots, p-1; \quad t = 1, 2, \cdots, T
\]  \hspace{1cm} (1)

Denote \( y^t = (y_1^t, \cdots, y_{p-1}^t)^\prime \), \( t = 1, 2, \cdots, T \). Obviously we have

\[
y_j^t \in (-\infty, +\infty), \quad j = 1, 2, \cdots, p-1; \quad \forall t = 1, 2, \cdots, T
\]

**Step 2** Using data of \( \{y_j^t, \quad t = 1, 2, 3, \cdots, T\} \), for \( j = 1, 2, \cdots, p-1 \), \( p-1 \) regression models can be preformed as follows:

\[
\hat{y}_j^t = f_j(t) + \varepsilon_j^t \quad j = 1, 2, \cdots, p - 1
\]  \hspace{1cm} (2)
**Step 3** Based on (2), the value of $y$ at the time $T+1$ could be calculated.

$$\hat{y}_{j}^{T+1} = f_{j}(T+l) \quad j = 1, 2, \cdots, p - 1 \quad (3)$$

**Step 4** Finally the predictive value of $X$ at the time $T+l$ can be obtained according to the equations (4) and (5) as follows:

$$x_{p}^{T+1} = \left\{ 1 / (1 + \sum_{j=1}^{p-1} e^{y_{j}^{T+1}}) \right\} \quad (4)$$

$$x_{j}^{T+1} = \left\{ e^{y_{j}^{T+1}} / (1 + \sum_{j=1}^{p-1} e^{y_{j}^{T+1}}) \right\} \quad , j = 1, 2, \cdots, p - 1 \quad (5)$$
Aitchison (1986) provides three ways for solving this modeling limitation:

(a) Amalgamation of finer parts to remove zero components.
(b) Replacing the zeros by very small values.
(c) Treating the zero observations as outliers.

However, all of these ways are not efficient if many such zero observations are presented in the data set. And the logratio transformation will fail in such applications.
3. Predictive Modeling (2)  
—— Hyperspherical-Transformation Approach

Basic Idea:

We firstly investigate a set of 3-D compositional data vector as follows.

\[
X' = \left\{ (x'_1, x'_2, x'_3)' \in \mathbb{R}^3 \left| \sum_{j=1}^{3} x'_j = 1, 0 \leq x'_j < 1 \right. \right\}, \quad t = 1,2,\cdots, T
\]

A square root transformation:

\[
y'_j = \sqrt{x'_j}, \quad j = 1,2,3; \quad t = 1,2,\cdots, T
\]

Since

\[
\|y'\|^2 = \sum_{j=1}^{3} (y'_j)^2 = 1
\]

All the vectors of \( y' = (y'_1, y'_2, y'_3)' \), \( t = 1,2,\cdots, T \), are on the surface of a 3-dimensional sphere with radius 1 exactly.
Mapping vector \( y^t = (y_1^t, y_2^t, y_3^t) \in \mathbb{R}^3 \) (\( t = 1, 2, \cdots, T \)) from the Cartesian coordinate system to spherical coordinate system \( (r^t, \theta_2^t, \theta_3^t) \in \Theta^3 \), and considering the condition of \( (r^t)^2 = \|y^t\|^2 = 1 \), there are the following equations for the mapping of \( \mathbb{R}^3 \rightarrow \Theta^2 \)

\[
\begin{align*}
    y_1^t &= \sin \theta_2^t \sin \theta_3^t \\
    y_2^t &= \cos \theta_2^t \sin \theta_3^t \\
    y_3^t &= \cos \theta_3^t
\end{align*}
\]

Where \( 0 < \theta_j^t \leq \pi / 2 \), \( j = 2, 3 \)

Thus the dimension of the vector is reduced from 3 down to 2.

\( y_1^t, y_2^t, y_3^t \rightarrow \theta_2^t, \theta_3^t \)
**A Hyperspherical-Transformation Forecasting Model for Compositional Data**

H. Wang, Q. Liu, H. M.K. Mok, L.Fu, W. M. Tse
Accepted By EJOR

---

**Step 1** Taking a square root transformation to the component of the original compositional data

\[
y_j^t = \sqrt{x_j^t}, \quad j = 1,2,\cdots, p; \quad t = 1,2,\cdots, T
\]  

(1)

Denote \( y^t = (y_1^t, \cdots, y_p^t)' \), \( t = 1,2,\cdots, T \). Obviously we have

\[
\|y^t\|^2 = \sum_{j=1}^{p} (y_j^t)^2 = 1
\]

(2)

Evidently, the extreme point of vector \( y^t = (y_1^t, \cdots, y_p^t)' \in \mathbb{R}^p \) is on the surface of a \( p \)-dimensional hyper-sphere with radius 1 at any time \( t \).
Step 2 Mapping vector \( y^t = (y_1^t, \ldots, y_p^t) \in \mathbb{R}^p \) \( t = 1, 2, \ldots, T \), from the Cartesian coordinate system to hyperspherical coordinate system \( (r^t, \theta_2^t, \ldots, \theta_p^t) \in \Theta^p \). Note the condition of

\[
(r^t)^2 = \|y^t\|^2 = 1
\]

there are the equations for the mapping of \( \mathbb{R}^p \to \Theta^{p-1} \) as follows,

\[
\begin{align*}
y_1^t &= \sin \theta_2^t \sin \theta_3^t \sin \theta_4^t \cdots \sin \theta_p^t \\
y_2^t &= \cos \theta_2^t \sin \theta_3^t \sin \theta_4^t \cdots \sin \theta_p^t \\
y_3^t &= \cos \theta_3^t \sin \theta_4 \cdots \sin \theta_p^t \\
&\vdots \\
y_{p-2}^t &= \cos \theta_{p-2}^t \sin \theta_{p-1}^t \sin \theta_p^t \\
y_{p-1}^t &= \cos \theta_{p-1}^t \sin \theta_p^t \\
y_p^t &= \cos \theta_p^t
\end{align*}
\]

where \( 0 < \theta_j^t \leq \pi / 2 \), \( j = 2, 3, \ldots, p \)
Step 3  In the transforms of Eq.(2) and (3), the dimension of compositional data is reduced from \( p \) down to \( p-1 \). The redundant of DoF has been deleted. According to Eq.(3), a inverse Transformation is applied to compute as follows,

\[
\begin{align*}
\theta'_p &= \arccos y'_p \\
\theta'_{p-1} &= \arccos \left( \frac{y'_{p-1}}{\sin \theta'_p} \right) \\
\theta'_{p-2} &= \arccos \left( \frac{y'_{p-2}}{\sin \theta'_p \sin \theta'_{p-1}} \right) \\
&\quad \vdots \\
\theta'_2 &= \arccos \left( \frac{y'_2}{\sin \theta'_p \sin \theta'_p \sin \theta'_3} \right) \\
\end{align*}
\]  

\[ \theta'_t = \arccos y'_t \]  

\[ t = 1,2,\cdots,T \]  

(4)
**Step 4** Based on the angle data \( \{\theta_j^t, t = 1, 2, 3, \ldots, T\} \), \( j = 2, 3, \ldots, p \), we can build \((p-1)\) models for each angle data series respectively, such as a regressive model

\[
\hat{\theta}_j^t = f_j(t) + \epsilon_j^t , \quad j = 2, 3, \ldots, p \tag{5}
\]

**Step 5** We use equation (5) to predict the angle at time \( T+l \) for different \( j \), as shown in Eq.(6)

\[
\hat{\theta}_j^{T+l} = f_j(T+l) , \quad j = 2, 3, \ldots, p \tag{6}
\]
Step 6  Computing the predicted value of $\hat{y}^{T+l} = (\hat{y}_1^{T+l}, \ldots, \hat{y}_p^{T+l})$ by using of Eq.(3). Because Eq.(3) is derived from condition of $(r^t)^2 = \|y^t\|^2 \equiv 1$, evidently we have

$$\sum_{j=1}^{p} (y_j^{T+l})^2 = 1 \tag{6}$$

Step 7  Finally, the predicted value of each component at time $T+l$ is obtained as

$$\hat{x}_j^{T+l} = (\hat{y}_j^{T+l})^2, \quad j = 1, 2, \ldots, p \tag{7}$$
Comparison with the Logratio Transformation

The assumption about the component value in HT approach is as follows,

\[ 0 \leq x_j^t < 1, \quad j = 1, 2, \ldots, p \]

The HT method will fail if and only if one component equals 1 and all other components equal 0.

Evidently, HT model compares favorably to Aitchison’s (1986) logratio transformation in this research mission.
4. Simulation Study

To verify the HP predictive modeling approach, a set of 4-D simulation data are generated by the following equations,

\[
\begin{align*}
    x_1 &= 0.0001t^3 - 0.0018t^2 + 0.0047t + 0.2646 + \text{normrnd}(0,0.005) \\
    x_2 &= 0.0002t^3 - 0.0042t^2 + 0.0030t + 0.4010 + \text{normrnd}(0,0.005) \\
    x_3 &= -0.0001t^3 + 0.0020t^2 - 0.0072t + 0.1436 + \text{normrnd}(0,0.005) \\
    x_4 &= 1 - x_1 - x_2 - x_3
\end{align*}
\]

Table 1 Data set generated for simulation study

<table>
<thead>
<tr>
<th>Time</th>
<th>Comp.</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>x_1</td>
<td></td>
<td>0.2656</td>
<td>0.2712</td>
<td>0.2712</td>
<td>0.2602</td>
<td>0.2503</td>
<td>0.2522</td>
<td>0.2327</td>
<td>0.2413</td>
<td>0.2370</td>
<td>0.2266</td>
</tr>
<tr>
<td>x_2</td>
<td></td>
<td>0.4034</td>
<td>0.3983</td>
<td>0.3716</td>
<td>0.3506</td>
<td>0.3431</td>
<td>0.3121</td>
<td>0.2845</td>
<td>0.2611</td>
<td>0.2304</td>
<td>0.2109</td>
</tr>
<tr>
<td>x_3</td>
<td></td>
<td>0.1424</td>
<td>0.1397</td>
<td>0.1372</td>
<td>0.1417</td>
<td>0.1411</td>
<td>0.1462</td>
<td>0.1518</td>
<td>0.1713</td>
<td>0.1698</td>
<td>0.1714</td>
</tr>
<tr>
<td>x_4</td>
<td></td>
<td>0.1886</td>
<td>0.1908</td>
<td>0.2201</td>
<td>0.2475</td>
<td>0.2655</td>
<td>0.2895</td>
<td>0.3309</td>
<td>0.3263</td>
<td>0.3629</td>
<td>0.3912</td>
</tr>
</tbody>
</table>
After Hyperspherical Transform we obtain the values of three angles $\theta_2^t, \theta_3^t, \theta_4^t$, $t = 1, 2, \ldots, 10$ as shown in Table 2.

<table>
<thead>
<tr>
<th>Angle</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_2^t$</td>
<td>0.6816</td>
<td>0.6899</td>
<td>0.7069</td>
<td>0.7112</td>
<td>0.7069</td>
<td>0.7323</td>
<td>0.7353</td>
<td>0.7656</td>
<td>0.7924</td>
<td>0.8033</td>
</tr>
<tr>
<td>$\theta_3^t$</td>
<td>1.1386</td>
<td>1.1422</td>
<td>1.1380</td>
<td>1.1220</td>
<td>1.1171</td>
<td>1.1000</td>
<td>1.0743</td>
<td>1.0423</td>
<td>1.0283</td>
<td>1.0116</td>
</tr>
<tr>
<td>$\theta_4^t$</td>
<td>1.1216</td>
<td>1.1187</td>
<td>1.0825</td>
<td>1.0501</td>
<td>1.0294</td>
<td>1.0027</td>
<td>0.9579</td>
<td>0.9628</td>
<td>0.9243</td>
<td>0.8951</td>
</tr>
</tbody>
</table>

Based on the computed angles in Table 2 three 3rd-order polynomial models are built for these angles respectively as follows,

$$
\begin{align*}
\theta_2^t &= 0.0001t^3 - 0.0004t^2 + 0.0053t + 0.6848 \\
\theta_3^t &= 0.0002t^3 - 0.0052t^2 + 0.0175t + 1.1240 \\
\theta_4^t &= 0.0003t^3 - 0.0060t^2 + 0.0110t + 1.1004
\end{align*}
$$
The fitted angles at time $t=1,2,\ldots,10$ and predicted angles at time $t=11, 12, \text{and} 13$ are plotted in Fig. 1.

Fig. 1 Angle values of $\theta_2$, $\theta_3$, and $\theta_4$
According to Eq.(3) and Eq.(7), the fitted components $x_1$, $x_2$, $x_3$ and $x_4$ at time $t=1,2,\ldots,10$ and predicted components at time $t=11, 12, \text{and} 13$ can be calculated, as shown in Fig. 2.

Fig. 2 Source data and predicted values in simulation

The simulation result demonstrates that the approach of HT predictive modeling is very successful.
5. Predictive modeling of the employment proportions in three industries of China

The observed data during 1990 and 1999 is used to investigate the trend of Chinese Three Industry Structure in last decade, as listed in the Table below.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary Industry</td>
<td>0.6010</td>
<td>0.5970</td>
<td>0.5850</td>
<td>0.5640</td>
<td>0.5430</td>
<td>0.5220</td>
<td>0.5050</td>
<td>0.4990</td>
<td>0.4980</td>
<td>0.5010</td>
</tr>
<tr>
<td>Secondary Industry</td>
<td>0.2140</td>
<td>0.2140</td>
<td>0.2170</td>
<td>0.2240</td>
<td>0.2270</td>
<td>0.2300</td>
<td>0.2350</td>
<td>0.2370</td>
<td>0.2350</td>
<td>0.2300</td>
</tr>
<tr>
<td>Tertiary Industry</td>
<td>0.1850</td>
<td>0.1890</td>
<td>0.1980</td>
<td>0.2120</td>
<td>0.2300</td>
<td>0.2480</td>
<td>0.2600</td>
<td>0.2640</td>
<td>0.2670</td>
<td>0.2690</td>
</tr>
</tbody>
</table>

*Data source: Chinese Statistic Yearbook, 1991~2000*
For $t=1990,\ldots, 1999$, the computed angles $\theta_2^t, \theta_3^t$ are obtained after hyperspherical mapping as listed in Table 5 and Fig.4.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta_2^t$</td>
<td>1.0328</td>
<td>1.0313</td>
<td>1.0238</td>
<td>1.0085</td>
<td>0.9848</td>
<td>0.9721</td>
<td>0.9674</td>
<td>0.9689</td>
<td>0.9753</td>
<td></td>
</tr>
<tr>
<td>$\theta_3^t$</td>
<td>1.1262</td>
<td>1.1210</td>
<td>1.1097</td>
<td>1.0923</td>
<td>1.0706</td>
<td>1.0495</td>
<td>1.0357</td>
<td>1.0312</td>
<td>1.0278</td>
<td>1.0255</td>
</tr>
</tbody>
</table>

Two 2nd order polynomial models are built for fitting the data in Table 5. 

\[
\theta_2^t = 0.0007t^2 - 0.0161t + 1.0572
\]

\[
\theta_3^t = 0.0008t^2 - 0.0214t + 1.1572
\]

Consequently, the predicted angles at $t=2000, 2001$ can be obtained

\[
\begin{align*}
\hat{\theta}_2 & = 0.5065 \\
\hat{\theta}_3 & = 0.9177 \\
\hat{\theta}_2 & = 0.5017 \\
\hat{\theta}_3 & = 0.9063
\end{align*}
\]
Based on the computed and predicted angles, we obtain the fitted and predicted values of $x_1, x_2, x_3$, as listed in Table 6 and Fig. 5.

**Table 6 Fitting values and Predicted values of Three Industries**

<table>
<thead>
<tr>
<th>Year</th>
<th>Fitting</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>0.6134</td>
<td>0.5924</td>
</tr>
<tr>
<td>$x_2$</td>
<td>0.2096</td>
<td>0.2157</td>
</tr>
<tr>
<td>$x_3$</td>
<td>0.1770</td>
<td>0.1918</td>
</tr>
<tr>
<td>$\Sigma$</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

Fig. 4 Computed angles and predicted angles

Fig. 5 Observed and predicted data of Three Industries
IV. Simple Linear Regression on Compositional Data

1. Issue

Construct a simple linear regression model when both independent and dependent variables are univariate compositional data;

\[
Y = \left(y_1, y_2, \ldots, y_q \right) \quad \sum_{j=1}^{q} y_j = 1 \quad , \quad 0 < y_j < 1
\]

\[
X = \left(x_1, x_2, \ldots, x_p \right) \quad \sum_{j=1}^{p} x_j = 1 \quad , \quad 0 < x_j < 1
\]

Characteristics:

- Regression with multiple dependent variables.
- The constrains of unit sum on both $Y$ and $X$ should be kept throughout modeling process
2. Simple Linear Regression Model of CD

Step 1. Adopt symmetrical logratio transformation

\[
u_j = \log \frac{y_j}{\sqrt[q]{\prod_{i=1}^{q} y_j}}, \; j = 1, 2, \ldots, q
\]

\[
v_j = \log \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_j}}, \; j = 1, 2, \ldots, p
\]

Problem on Regression:

The independent variables are completely correlated.
Step 2. Construct the PLS regression model.
\[ \hat{u}_j = \alpha_{j0} + \alpha_{j1} v_1 + \cdots + \alpha_{jp} v_p, \quad j = 1, 2, \ldots, q \]

Step 3. According to the models above, we get the predictive value of \( \hat{U} = (\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_q)' \)

Step 4. Adopt the inverse transformation of symmetrical logratio transformation, we finally obtain the predictive results as \( \hat{Y} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_q)' \)

\[
\hat{w}_j = \hat{u}_j - \hat{u}_q, \quad j = 1, 2, \ldots, q - 1 \\
\hat{y}_q = \left\{ \frac{1}{1 + \sum_{i=1}^{q-1} e^{\hat{w}_i}} \right\} \\
\hat{y}_j = e^{\hat{w}_j} \hat{y}_q, \quad j = 1, 2, \ldots, q - 1
\]
3. Case Study

Optimization and upgrade of industrial structure is critical to the economic development in an area; meanwhile, changes in productive and economic structures inevitably affect employment structure. The trend-lines of GDP and employment structures of Beijing’s three industries from 1990 to 2001 are shown in figure below which presents an obvious linear relationship between these two compositional data.
### Employment portion and GDP portion in Beijing

<table>
<thead>
<tr>
<th>Year</th>
<th>$Y_1$</th>
<th>$Y_2$</th>
<th>$Y_3$</th>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1990</td>
<td>14.46</td>
<td>44.91</td>
<td>40.63</td>
<td>8.76</td>
<td>52.39</td>
<td>38.85</td>
</tr>
<tr>
<td>1991</td>
<td>14.32</td>
<td>44.12</td>
<td>41.56</td>
<td>8.15</td>
<td>52.18</td>
<td>39.67</td>
</tr>
<tr>
<td>1992</td>
<td>13.01</td>
<td>43.37</td>
<td>43.62</td>
<td>6.87</td>
<td>48.78</td>
<td>44.35</td>
</tr>
<tr>
<td>1993</td>
<td>12.29</td>
<td>43.55</td>
<td>44.16</td>
<td>6.21</td>
<td>48.03</td>
<td>45.76</td>
</tr>
<tr>
<td>1994</td>
<td>11.02</td>
<td>40.98</td>
<td>48.01</td>
<td>6.81</td>
<td>46.15</td>
<td>47.04</td>
</tr>
<tr>
<td>1995</td>
<td>10.61</td>
<td>40.73</td>
<td>48.65</td>
<td>5.84</td>
<td>44.1</td>
<td>50.06</td>
</tr>
<tr>
<td>1996</td>
<td>10.98</td>
<td>39.40</td>
<td>49.62</td>
<td>5.17</td>
<td>42.28</td>
<td>52.55</td>
</tr>
<tr>
<td>1997</td>
<td>10.69</td>
<td>39.28</td>
<td>50.03</td>
<td>4.69</td>
<td>40.8</td>
<td>54.51</td>
</tr>
<tr>
<td>1998</td>
<td>11.49</td>
<td>36.32</td>
<td>52.19</td>
<td>4.3</td>
<td>39.12</td>
<td>56.58</td>
</tr>
<tr>
<td>1999</td>
<td>12.04</td>
<td>34.95</td>
<td>53.01</td>
<td>4</td>
<td>38.7</td>
<td>57.3</td>
</tr>
<tr>
<td>2000</td>
<td>11.77</td>
<td>33.62</td>
<td>54.61</td>
<td>3.7</td>
<td>38</td>
<td>58.5</td>
</tr>
<tr>
<td>2001</td>
<td>11.31</td>
<td>34.33</td>
<td>54.36</td>
<td>3.3</td>
<td>36.2</td>
<td>60.5</td>
</tr>
</tbody>
</table>
There is a strong correlation relationship between the $1_{th}$ PLS components $t_1$ and $u_1$. 
With PLSR, according to cross-validation, we select two component from $X$, which can explain 81.4% of explanatory variables and 88% of response, with the following models. And the Goodness of Fit of these three models are 81.3%, 88.5% and 97.0% respectively.

\[
Y_1 = 16.5975 - 2.86001X_1 + 1.51719X_2 - 2.00526X_3 \\
Y_2 = 28.448 + 0.725227X_1 + 0.14436X_2 - 0.0768685X_3 \\
Y_3 = 28.1363 + 0.460484X_1 - 0.658354X_2 + 0.780977X_3
\]
1. Issue

Construct a simple linear regression model when independent variable are multivariate compositional data.

◆ Following problems should be concerned:

(1) it is a regression model with multiple compositional data as independent variables;
(2) the sum to unity constraint should be satisfied in each compositional data throughout the modeling.
(3) Independent variables are totally collinear.
(4) The hierarchy relationship within the compositional data should be considered.
## Hierarchy Relationship Within Compositional Data

<table>
<thead>
<tr>
<th>Y</th>
<th>X₁</th>
<th>X₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>Employment Proportion</td>
<td>GDP Proportion</td>
<td>Investment Proportion</td>
</tr>
<tr>
<td>emp1</td>
<td>emp2</td>
<td>emp3</td>
</tr>
</tbody>
</table>

\[
\hat{Y} = f(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})
\]

\[
\hat{\eta} = f(\xi_1, \xi_2)
\]

with \( \eta = h(y_{11}, y_{12}, y_{13}) \)

\( \xi_1 = g_1(x_{11}, x_{12}, x_{13}) \)

\( \xi_2 = g_2(x_{21}, x_{22}, x_{23}) \)
2. Multivariate Compositional Data Modeling (MCDM)

To solve the problem of hierarchy relationship of independent variables, PLS Path modeling method is applied in MCDM.

**Modeling Procedures:**

**Step1** apply symmetrical logratio transformation to the different compositional data of independent variables and dependent variable.

**Step2** apply PLS Path modeling to the transformed variables, and extract the latent variables of each compositional data.

**Step3** build multivariate regression model of the LVs and analyze their relationship.

\[
\begin{align*}
&\begin{bmatrix} X_1, \ldots, X_m \end{bmatrix} \\
&Y \\
&\text{symmetrical logratio transformation} \\
&\begin{bmatrix} \tilde{X}_1, \ldots, \tilde{X}_m \end{bmatrix} \\
&\tilde{Y} \\
&\text{PLS Path Modeling} \\
&\text{LVs} \\
\end{align*}
\]

Regression Modeling
For the Forecast Purpose:

- Take symmetrical logratio transformation to the dependent variable sample $X^{(0)}$;
- Put the transformed variables $\tilde{X}^{(0)}$ in the built model, so that the corresponding transformed dependent variables $\tilde{Y}^{(0)}$ can be calculated;
- Applying the inverse transformation, the values of dependent variables $\hat{Y}^{(0)}$ can be worked out eventually.
3. Application

MCDM is applied to build regression model between the proportion investment, GDP and total employment in Beijing’s three industries depends on the data from 1990 to 2003.

Table 1 defines latent variables (LVs) and manifest variables (MVs) in the model.

<table>
<thead>
<tr>
<th>LVs</th>
<th>Investment Proportion (inv., $\xi_1$)</th>
<th>GDP Proportion (GDP, $\xi_2$)</th>
<th>Employment Proportion (emp., $\xi_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVs</td>
<td>inv$_1$</td>
<td>GDP$_1$</td>
<td>emp$_1$</td>
</tr>
<tr>
<td></td>
<td>inv$_2$</td>
<td>GDP$_2$</td>
<td>emp$_2$</td>
</tr>
<tr>
<td></td>
<td>inv$_3$</td>
<td>GDP$_3$</td>
<td>emp$_3$</td>
</tr>
</tbody>
</table>
Table2 is the primary **inner design matrix** and fig.1 is the corresponding **path model graph**.

**Table2. Inner design matrix of the structural model (1)**

<table>
<thead>
<tr>
<th></th>
<th>inv.</th>
<th>GDP</th>
<th>emp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>inv.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GDP</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emp.</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Fig.1. Path model of investment, GDP and employment**

Apply PLS Path method to the transformed data and obtain the path coefficients of LVs in table3.

**Table3. Path coefficients of LVs (1)**

<table>
<thead>
<tr>
<th></th>
<th>inv.</th>
<th>GDP</th>
<th>emp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>inv.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GDP</td>
<td>0.872</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emp.</td>
<td>0.004</td>
<td>0.948</td>
<td>0</td>
</tr>
</tbody>
</table>

Because the direct effect of inv. to emp., 0.004, is too small, this path can be removed.
The modified inner design matrix and its corresponding path coefficients of LVs are listed in table4 and table5.

Table 4. Inner design matrix (2)

<table>
<thead>
<tr>
<th></th>
<th>inv.</th>
<th>GDP</th>
<th>emp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>inv.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GDP</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emp.</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 5. Path coefficients of LVs (2)

<table>
<thead>
<tr>
<th></th>
<th>inv.</th>
<th>GDP</th>
<th>emp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>inv.</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GDP</td>
<td>0.872</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>emp.</td>
<td>0</td>
<td>0.951</td>
<td>0</td>
</tr>
</tbody>
</table>

The indirect effect of inv. to emp. through GDP is \(0.872 \times 0.951 = 0.829\).

Fig. 5 shows that inv. is important in the prediction of GDP, with \(R^2\) 75.99%, and GDP to emp. 90.51%.
PLS Regression is applied to describe the relationships of each LV and its corresponding MVs.

\[
\text{inv.} = 0.4318 \text{inv}_1 - 0.1651 \text{inv}_2 - 0.4859 \text{inv}_3 \\
\text{GDP} = 0.4327 \text{GDP}_1 + 0.2479 \text{GDP}_2 - 0.4414 \text{GDP}_3 \\
\text{emp.} = 0.2815 \text{emp}_1 + 0.4182 \text{emp}_2 - 0.4516 \text{emp}_3
\]

The relationship between LVs can be described by the following polynomial model.

\[
\text{emp.} = -0.1812 + 0.1812 (\text{inv.})^2 + 0.8686 \text{GDP}
\]

The expressions display the impact extent of the investment and GDP to the employment as well as their quantitative relationships.
Predictive Modeling for Gini Coefficient

H. Wang, W Long, Q. Liu,, Preceding of the 5th International Conference on Management, Macao, 2004
Lorenz Curve and Gini Coefficient are generally used to depict the disproportion of the income distribution. As effective statistical tools for variable distribution analysis, these two approaches have been widely applied to economics.

As an effective statistical tool for variable distribution analysis, Lorenz Curve and Gini Coefficient are not only used in the income distribution study, but also in the analysis of many distribution problems, such as the imbalance of market, recourses, etc. Therefore, they are also called the generalized approaches for distribution analysis.
Generally, Lorenz Curve is a concave curve with the numerical values in the range of $[0,1]$, as shown in Fig. 1, the axis $P$ represent the accumulative rates of the population, and the axis $I$ represent the accumulative rates of income of the corresponding population group.
Gini Coefficient

Gini Coefficient, defined by G. Gini, can describe the extent of the income disproportion.

In Fig. below, we suppose that $S_A$ represents the area between Lorenz Curve and “the absolute equality line”, and $S_B$ the area between Lorenz Curve and the horizontal axis. Therefore we can define the Gini Coefficient as:

$$G = \frac{S_A}{S_A + S_B} = 2S_A$$

Since

$$S_A + S_B = \frac{1}{2}$$
2. Predictive modeling for Lorenz Curve and Gini Coefficient

(1) Predictive Modeling of Compositional Data

The prediction of Gini Coefficient should be based on the prediction of population and income distribution.

And the predictive modeling of both population and income distribution can be built respectively by use of the predictive modeling of compositional data.
(2) Predictive Modeling of Gini Coefficient

Step 1. By use of predictive modeling of compositional data, we predict the future population distribution and income distribution.

Step 2. Regress an equation fitting for Lorenz Curve.

Step 3. Make integration under the curve in the interval [0,1] so as to get the area $S_B$.

Step 4. Since $S_A + S_B = 1/2$, thus

$$S_A = (1/2 - S_B)$$

Hereby Gini Coefficient can be achieved as follows:

$$G = 2S_A$$
3. Case study: the imbalance analysis on the GDP in China

The imbalance of economic development in the different regions is one of the remarkable issues in China. According to the level of Per capita GDP, all the 30 provinces/municipalities/autonomous regions in China can be divided into 5 groups shown as in Table below.

<table>
<thead>
<tr>
<th>Name</th>
<th>Regions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group No.1</td>
<td>Shanghai, Beijing, Tianjin, Zhejiang, Guangdong, Jiangsu</td>
</tr>
<tr>
<td>Group No.2</td>
<td>Fujian, Liaoning, Shandong, Heilongjiang, Hebei</td>
</tr>
<tr>
<td>Group No.3</td>
<td>Sinkiang, Hubei, Jilin, Hainan, Inner Mongolia, Hunan</td>
</tr>
<tr>
<td>Group No.4</td>
<td>Henan, Qinghai, Shanxi, Anhui, Jiangxi</td>
</tr>
<tr>
<td>Group No.5</td>
<td>Tibet, Sichuan, Shaanxi, Ningxia, Yunnan, Guangxi, Gansu, Guizhou</td>
</tr>
</tbody>
</table>

Group No.1, with 19% of the population of the whole country, creates more than 34% of GDP, whereas Group No.5, with 30% of the population, only creates approximately 14% of GDP.
The regions of Group 1 and Group 2, which are mainly constituted of east provinces and cities along the coast, only corresponding to 1/6 of whole mainland regions in China. The total population of these region covers 40% of the whole country. On the other hand, GDP of this area occupies 60% of the whole country in 2001.
The Lorenz curves of Population vs GDP in 1991 and 2001 can be drawn by use of the chronological values of population and GDP of different regions corresponding to the five Groups.

And the Gini coefficient of Population vs GDP in 1991 is calculated as 0.199 while it increases to 0.256 in 2001 in China.
The predictive models of the proportions of population and GDP in five regions in China can be built according to logratio approach. Then the fitting values and the predicted values are shown in Figures below.

![Fig. 3 Fitting values of population proportion of five regions](image)

![Fig. 4 Fitting values of proportion GDP of five regions](image)

Obviously, the proportions variety of the population among all the regions is very slight. But on the other hand, the variety of GDP among different regions has a wide range. Especially, the GDP of the first region shows an evident increasing tendency.
Using the predicted data in the table2, the Lorenz curve of GDP in 2005 can be plotted as shown in Figure 5.

<table>
<thead>
<tr>
<th>Group No.1</th>
<th>Group No.2</th>
<th>Group No.3</th>
<th>Group No.4</th>
<th>Group No.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population rate</td>
<td>0.1874</td>
<td>0.2121</td>
<td>0.1602</td>
<td>0.1886</td>
</tr>
<tr>
<td>GDP rate</td>
<td>0.3557</td>
<td>0.2606</td>
<td>0.1287</td>
<td>0.1257</td>
</tr>
</tbody>
</table>

Correspondingly, Gini coefficient of GDP in 2005 is calculated as 0.276.

It is seen that the imbalance of different regions has been enlarged in China. And this situation may become more grave in the future years if there will not be efficient policies’ intervention from the Chinese government.


