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.

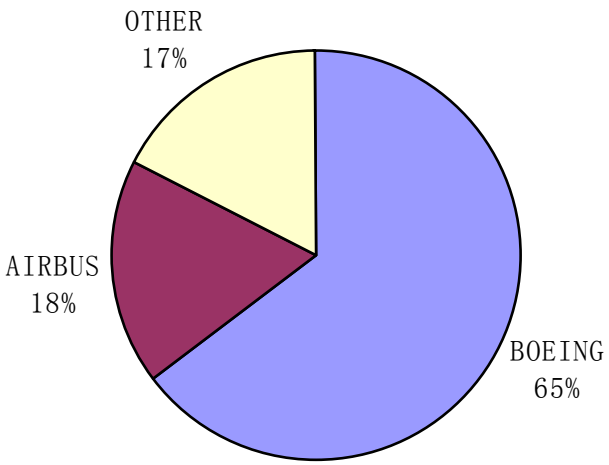
For Example:

- 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

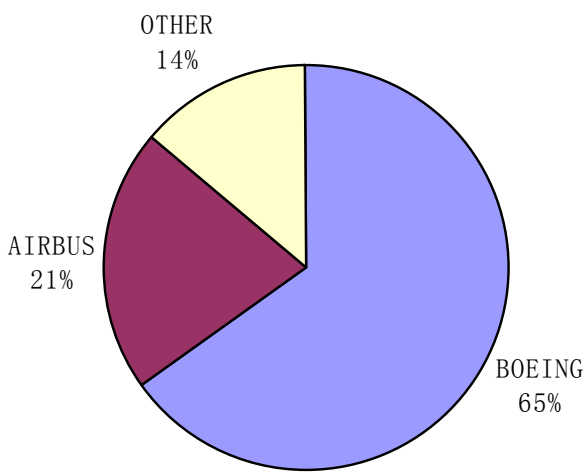
- 1985： AIRBUS sold the first civil airplane to China
- 1995： There were only 29 airplanes of AIRBUS served in China
- 2005：1/6 airplanes sold in AIRBUS was destined to China (65 airplanes)
- In 2011, China Civil Airline will need about 1600 airplanes. And market share of AIRBUS would arrive at 50% in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

2000



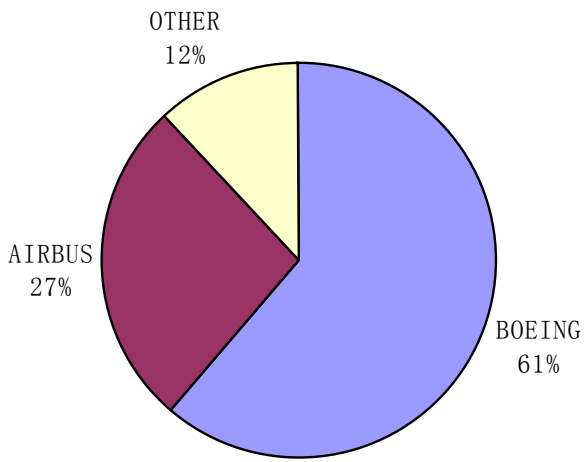
527 airplanes

2003



Pie Chat

2005



863 airplanes

2. Definition of Compositional Data

Compositional Data:

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

$$\mathbf{X} = (x_1, x_2, \cdots, x_p)' \in \mathbf{R}^p$$

which is subject to the constraints below,

$$\sum_{j=1}^{p} x_j = 1 \ , \qquad x_j \geq 0$$

Hereafter we often call x_j, j=1,2,…,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.

Motivation of the Course

(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.

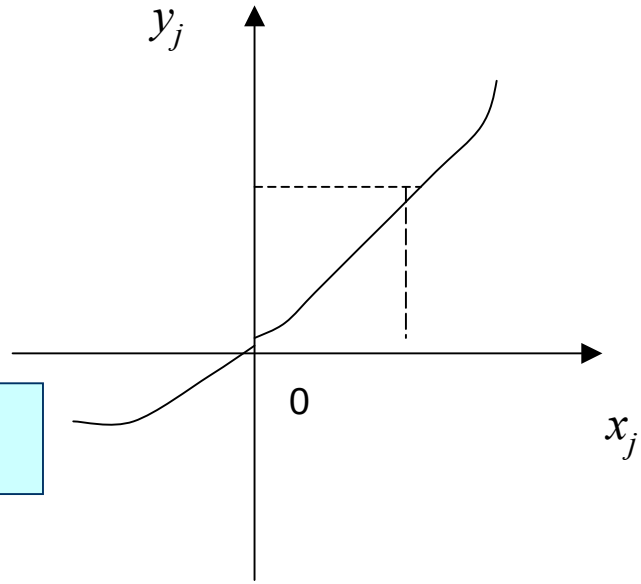
(2) Regression Model

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):

$$y_j = \ln\left(x_j / x_p\right),\ j = 1, 2, .., p-1$$



Reason: $\mathbf{X} = (x_1, x_2, \cdots, x_p)' \in \mathbf{R}^p$

$0 \le x_j \le 1, \quad j = 1, 2, \cdots, p$ ← Delete this constrains

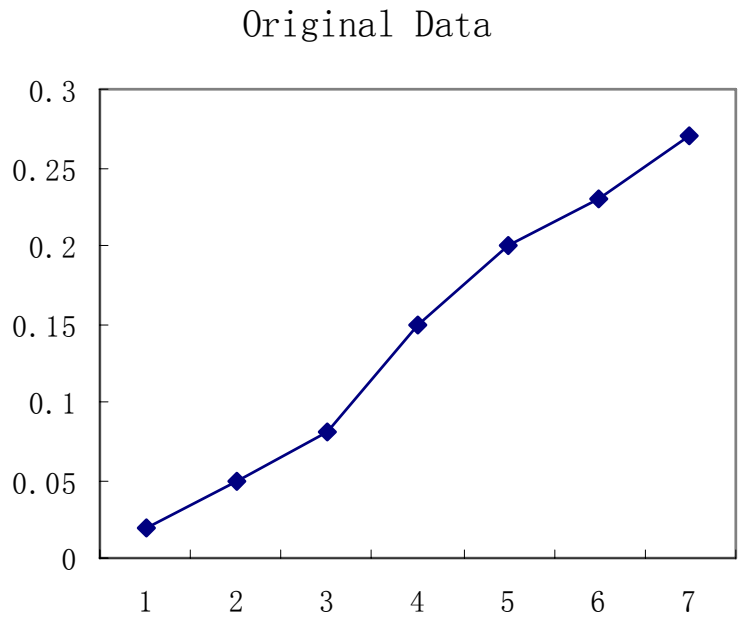
Let $u_j = x_j / x_p$

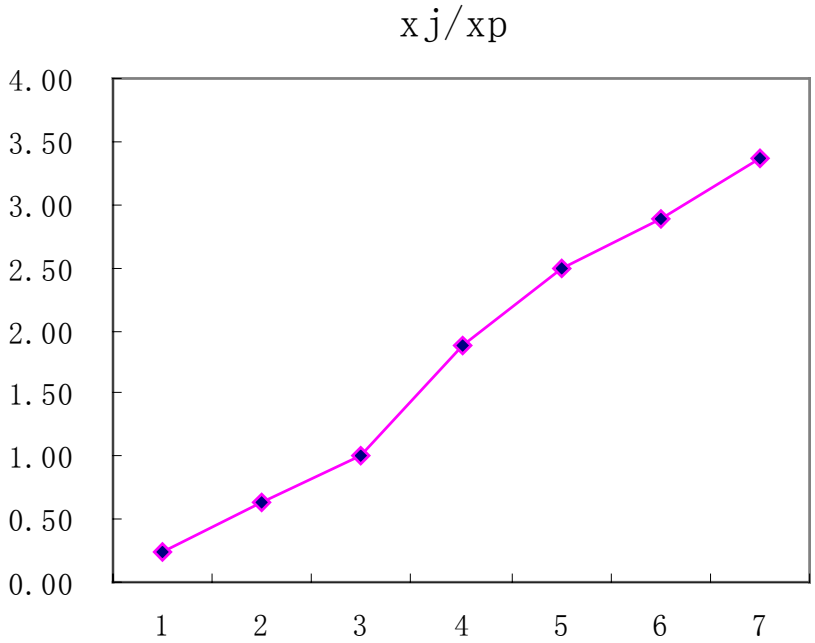
We get: $u_1 = x_1 / x_p,\ u_2 = x_2 / x_p,\ \cdots, u_{p-1} = x_{p-1} / x_p,\ u_p = 1$

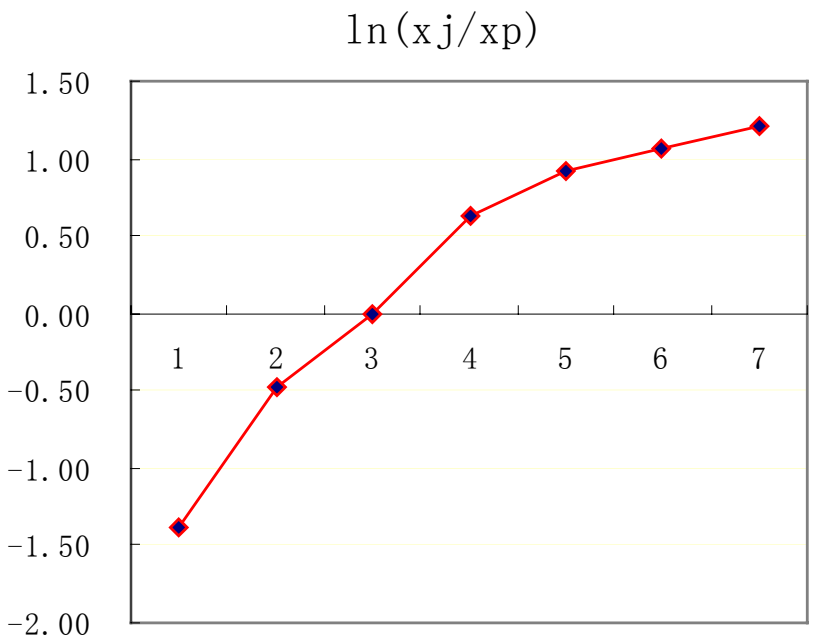
And $u_j \in [0, +\infty)$

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

x_j	x_j/x_p	$\ln (x_j/x_p)$
0.02	0.25	-1.39
0.05	0.63	-0.47
0.15	1.88	0.63
0.2	2.50	0.92
0.23	2.88	1.06
0.27	3.38	1.22
0.08	1.00	0.00







Advantage of logratio transformation

- Firstly, it is easier to choose a function fitting for the model as the value of y_j^t 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 $\boldsymbol{X}$ follows additive logistic normal distribution, the transformed vector $\boldsymbol{Y}$ will follow the normal distribution.

Inverse Transformation

Since: $y_j = \ln(x_j / x_p),\ j = 1, 2 \cdots, p-1$

$$x_j = x_p e^{y_j},\quad j = 1, 2 \cdots, p-1$$

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:

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

$$x_j = \left\{ e^{y_j} / (1 + \sum_{j=1}^{p-1} e^{y_j}) \right\},\ j = 1, 2, \cdots, p-1$$

Inconvenient

(1) Short of exploitability to the modeling results

$$y_j = \ln\left(x_j / x_p\right), j = 1, 2, .., p-1$$

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)

$$0 < x_j^t < 1, \quad j = 1, 2, \cdots, p$$

2. Symmetrical Logratio Transformation:

(Aitchison, 1986)

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

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

$$x_j = e^{w_j} x_p, \quad j = 1,2,\cdots,p$$

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

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

$$x_j = \frac{e^{w_j}}{\sum_{j=1}^{p-1} e^{w_j} + 1}, \quad j = 1,2,\cdots,p-1$$

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,

$$X^t = \left\{ (x_1^t, \cdots, x_p^t)' \in \mathbf{R}^p \left| \sum_{j=1}^{t} x_j^t = 1, 0 \le x_j^t \le 1 \right. \right\}, \qquad t = 1,2,\cdots,T$$

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

$$X^{T+l} = \left\{ (x_1^{T+l}, \cdots, x_p^{T+l})' \in \mathbf{R}^p \left| \sum_{j=1}^{p} x_j^{T+l} = 1,\ 0 \le x_j^{T+l} \le 1 \right. \right\}$$

The changes of three industrial structures happened in 1952,1985 and 2000

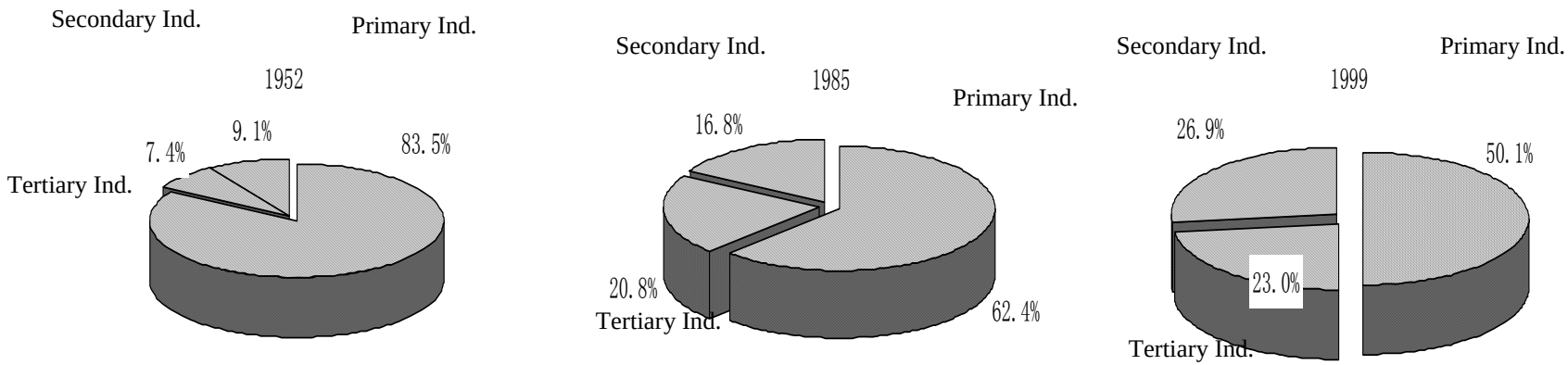


Fig. 1 Ratios of GDP of three industries in China

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,…,p, a model is built according to the data record. Then the model is used for prediction of future data.

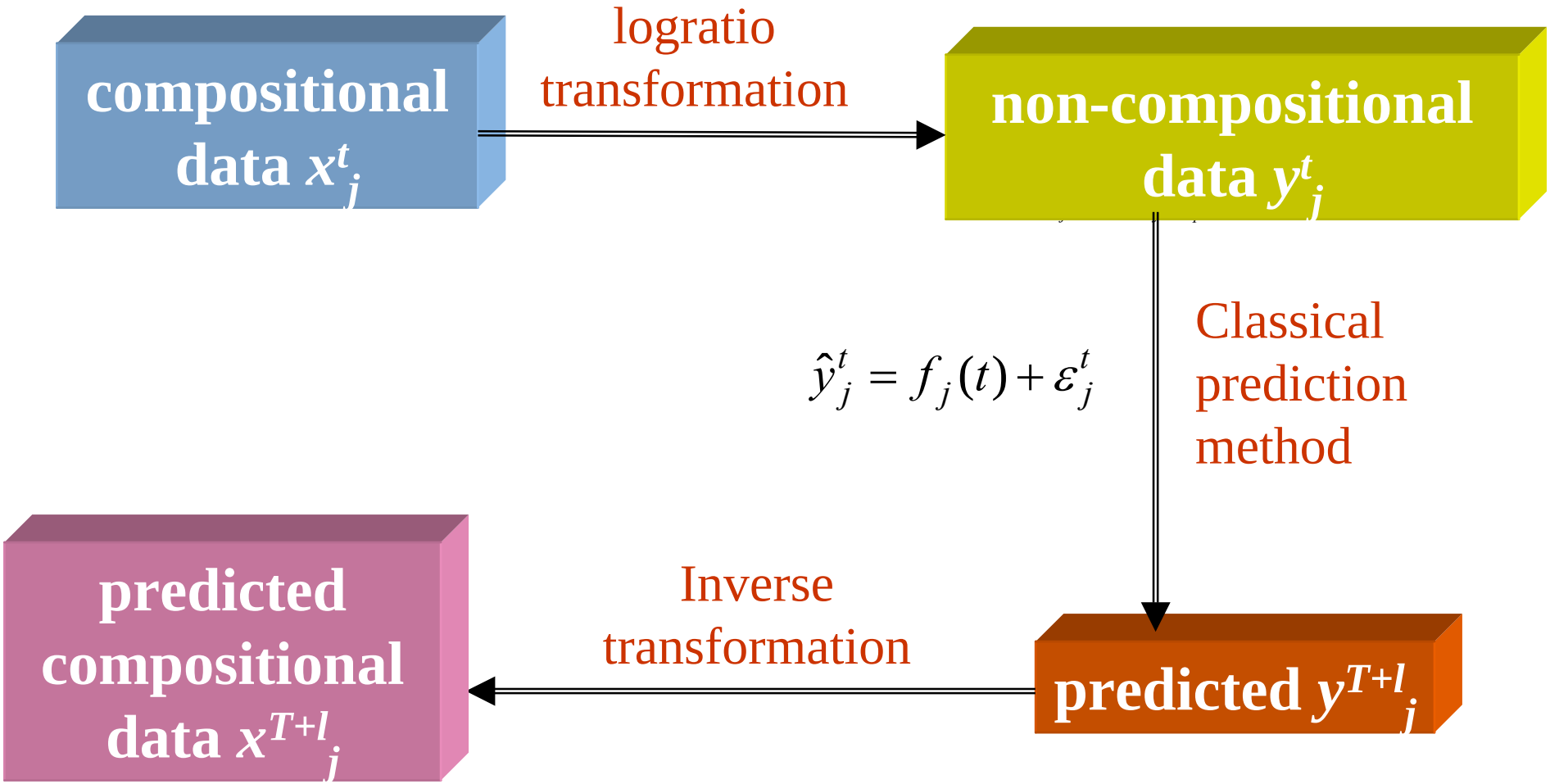
$$x_j^t, \quad t = 1,2,\cdots,T \quad \rightarrow \quad x_j^{T+l}$$

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



3. Predictive Modeling (1)

—— by use of logratio transformation

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

$$y_j^t = \ln(x_j^t / x_p^t)\ , j = 1,2,\cdots, p-1;\ t = 1,2,\cdots, T \qquad (1)$$

Denote $y^t = \left(y_1^t, \cdots, y_{p-1}^t\right)'$, $\mathrm{t} = 1,2,\cdots,\mathrm{T}$. Obviously we have

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

Step2 Using data of $\left\{y_j^t\ ,\ t = 1,2,3,\cdots, T\right\}$, for $j = 1,2,\cdots, p-1$,

（p-1） regression models can be preformed as follows:

$$\hat{y}_j^t = f_j(t) + \varepsilon_j^t \quad , j = 1,2,\cdots, p\text{-}1 \qquad (2)$$

Step 3 Based on (2), the value of y at the time T+l could be calculated..

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

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:

$$x_p^{T+l} = \left\{ 1/(1+\sum_{j=1}^{p-1} e^{y_j^{T+l}}) \right\} \qquad (4)$$

$$x_j^{T+l} = \left\{ e^{y_j^{T+l}} /(1+\sum_{j=1}^{p-1} e^{y_j^{T+l}}) \right\} \quad , j = 1,2,\cdots, p-1 \qquad (5)$$

Limitation:

$$0 < x_j^t < 1, \quad j = 1,2,\cdots,p$$

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.

$$X^t = \left\{ (x_1^t, x_2^t, x_3^t)' \in \mathbf{R}^3 \,\middle|\, \sum_{j=1}^{3} x_j^t = 1, 0 \le x_j^t < 1 \right\}, \qquad t = 1,2,\cdots,T$$

A square root transformation:

$$y_j^t = \sqrt{x_j^t}, \; j = 1,2,3; \;\; t = 1,2,\cdots,T$$

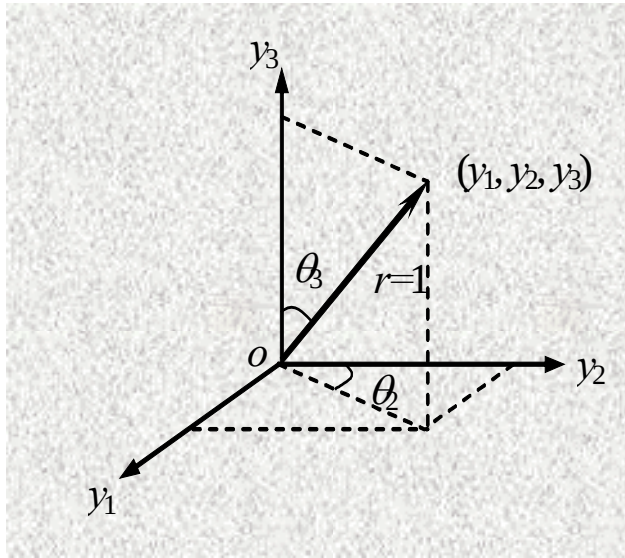
Since

$$\left\| y^t \right\|^2 = \sum_{j=1}^{3} (y_j^t)^2 = 1$$

All the vectors of $y^t = \left(y_1^t, y_2^t, y_3^t\right)', \quad \mathrm{t} = 1,2,\cdots,\mathrm{T}$ **, are on the surface of a 3-dimensional sphere with radius 1 exactly.**

Mapping vector $y^t = \left(y_1^t, y_2^t, y_3^t\right)' \in \mathbf{R}^3$ **(** $t = 1,2,\cdots,T$ **) from the Cartesian coordinate system to spherical coordinate system** $\left(r^t, \theta_2^t,, \theta_3^t\right)' \in \Theta^3$ **, and considering the condition of** $(r^t)^2 = \left\|y^t\right\|^2 \equiv 1$ **, there are the following equations for the mapping of** $\mathbf{R}^3 \to \Theta^2$

$$\begin{cases} y_1^t = \sin\theta_2^t \sin\theta_3^t \\ y_2^t = \cos\theta_2^t \sin\theta_3^t \\ y_3^t = \cos\theta_3^t \end{cases}$$



Where $0 < \theta_j^t \le \pi/2 \ , \ j = 2,3$

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

$$y_1^t, y_2^t, y_3^t \longrightarrow \theta_2^t, \theta_3^t$$

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

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

Denote $y^t = \left(y_1^t,\cdots,y_p^t\right)'$, $\mathrm{t} = 1,2,\cdots,\mathrm{T}$. Obviously we have

$$\left\|y^t\right\|^2 = \sum_{j=1}^{p}(y_j^t)^2 = 1 \tag{2}$$

Evidently, the extreme point of vector $y^t = \left(y_1^t,\cdots,y_p^t\right)' \in \mathbf{R}^p$, is on the surface of a *p*-dimensional hyper-sphere with radius 1 at any time *t*.

Step 2 Mapping vector $y^t = \left(y_1^t, \cdots, y_p^t\right)' \in \mathbf{R}^p \quad t = 1,2,\cdots,T$, from the Cartesian coordinate system to hyperspherical coordinate system $\left(r^t, \theta_2^t, \cdots, \theta_p^t\right)' \in \Theta^p$. Note the condition of

$$(r^t)^2 = \left\| y^t \right\|^2 \equiv 1$$

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

$$\begin{cases} y_1^t = \sin\theta_2^t \sin\theta_3^t \sin\theta_4^t \cdots \sin\theta_p^t \\ y_2^t = \cos\theta_2^t \sin\theta_3^t \sin\theta_4^t \cdots \sin\theta_p^t \\ y_3^t = \cos\theta_3^t \sin\theta_4^t \cdots \sin\theta_p^t \\ \quad \vdots \\ y_{p-2}^t = \cos\theta_{p-2}^t \sin\theta_{p-1}^t \sin\theta_p^t \\ y_{p-1}^t = \cos\theta_{p-1}^t \sin\theta_p^t \\ y_p^t = \cos\theta_p^t \end{cases} \tag{3}$$

where $0 < \theta_j^t \le \pi/2$, $j = 2,3,\cdots,p$

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,

$$\begin{cases} \theta_p^t = \arccos y_p^t \\ \theta_{p-1}^t = \arccos\left(\dfrac{y_{p-1}^t}{\sin\theta_p^t} \right) \\ \theta_{p-2}^t = \arccos\left(\dfrac{y_{p-2}^t}{\sin\theta_p^t \sin\theta_{p-1}^t} \right) \\ \quad \vdots \\ \theta_2^t = \arccos\left(\dfrac{y_2^t}{\sin\theta_p^t \sin\theta_{p-1}^t \cdots \sin\theta_3^t} \right) \end{cases} \qquad t = 1,2,\cdots,T \qquad (4)$$

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

$$\hat{\theta}_j^t = f_j(t) + \varepsilon_j^t \quad , j=2,3,\cdots,p \qquad (5)$$

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

$$\hat{\theta}_j^{T+l} = f_j(T+l) \quad , j=2,3,\cdots,p \qquad (6)$$

Step 6 Computing the predicted value of $\hat{y}^{T+l} = \left(\hat{y}_1^{T+l}, \cdots, \hat{y}_p^{T+l}\right)$ by using of Eq.(3). Because Eq.(3) is derived from condition of $(r^t)^2 = \left\|y^t\right\|^2 \equiv 1$, evidently we have

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

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

$$\hat{x}_j^{T+l} = \left(\hat{y}_j^{T+l}\right)^2, \quad j = 1,2,\cdots,p \qquad (7)$$

Comparison with the Logratio Transformation

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

$$0 \leq x_j^t < 1, \quad j = 1,2,\cdots, p$$

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 ,

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

Table 1 Data set generated for simulation study

Time / Comp.	1	2	3	4	5	6	7	8	9	10
x_1	0.2656	0.2712	0.2712	0.2602	0.2503	0.2522	0.2327	0.2413	0.2370	0.2266
x_2	0.4034	0.3983	0.3716	0.3506	0.3431	0.3121	0.2845	0.2611	0.2304	0.2109
x_3	0.1424	0.1397	0.1372	0.1417	0.1411	0.1462	0.1518	0.1713	0.1698	0.1714
x_4	0.1886	0.1908	0.2201	0.2475	0.2655	0.2895	0.3309	0.3263	0.3629	0.3912

After Hyperspherical Transform we obtain the values of three angles $\theta_2^t \quad \theta_3^t \quad \theta_4^t$, $t = 1,2,\cdots,10$ as shown in Table 2.

Table 2 Computed angles

Angle \ Time	1	2	3	4	5	6	7	8	9	10
θ_2^t	0.6816	0.6899	0.7069	0.7112	0.7069	0.7323	0.7353	0.7656	0.7924	0.8033
θ_3^t	1.1386	1.1422	1.1380	1.1220	1.1171	1.1000	1.0743	1.0423	1.0283	1.0116
θ_4^t	1.1216	1.1187	1.0825	1.0501	1.0294	1.0027	0.9579	0.9628	0.9243	0.8951

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

$$\theta_2^t = 0.0001t^3 - 0.0004t^2 + 0.0053t + 0.6848$$

$$\theta_3^t = 0.0002t^3 - 0.0052t^2 + 0.0175t + 1.1240$$

$$\theta_4^t = 0.0003t^3 - 0.0060t^2 + 0.0110t + 1.1004$$

The fitted angles at time t=1,2,…,10 and predicted angles at time t=11, 12, and 13 are plotted in Fig. 1.

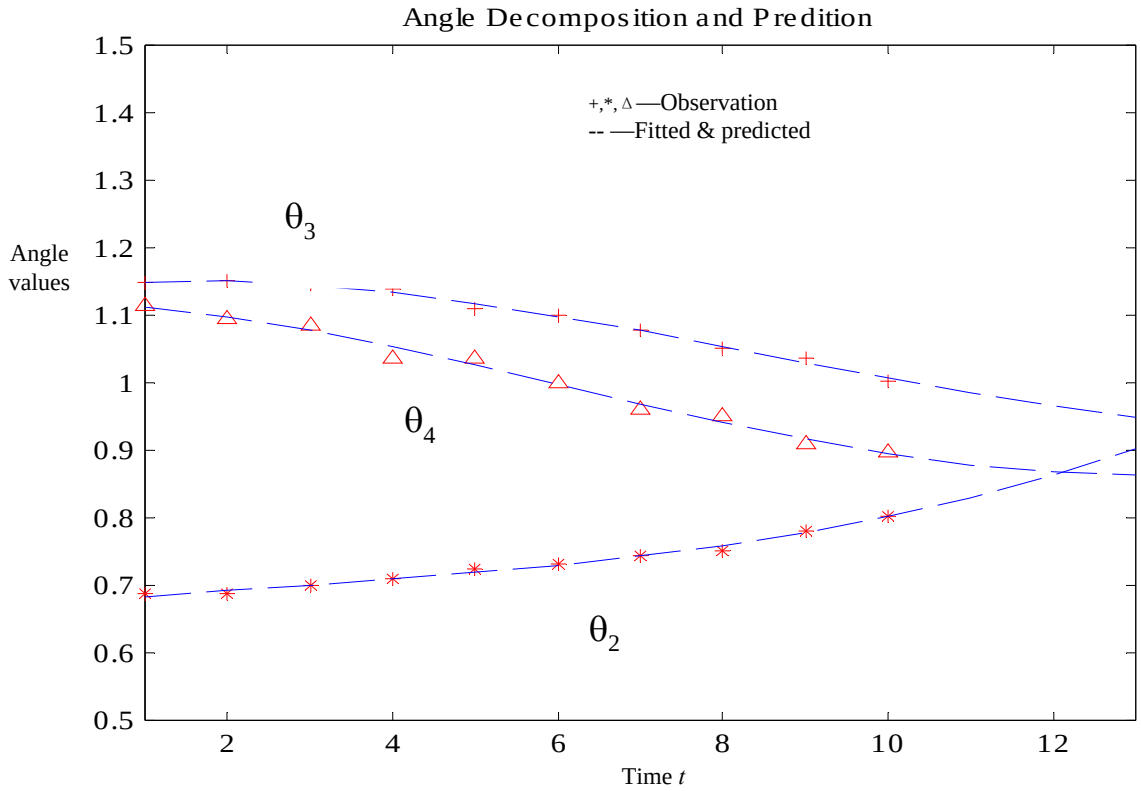


Fig. 1 Angle values of , θ_2^t θ_3^t and θ_4^t

According to Eq.(3) and Eq.(7), the fitted components x_1, x_2, x_3 and x_4 at time t=1,2,…,10 and predicted components at time t=11, 12, and 13 can be calculated, as shown in Fig. 2.

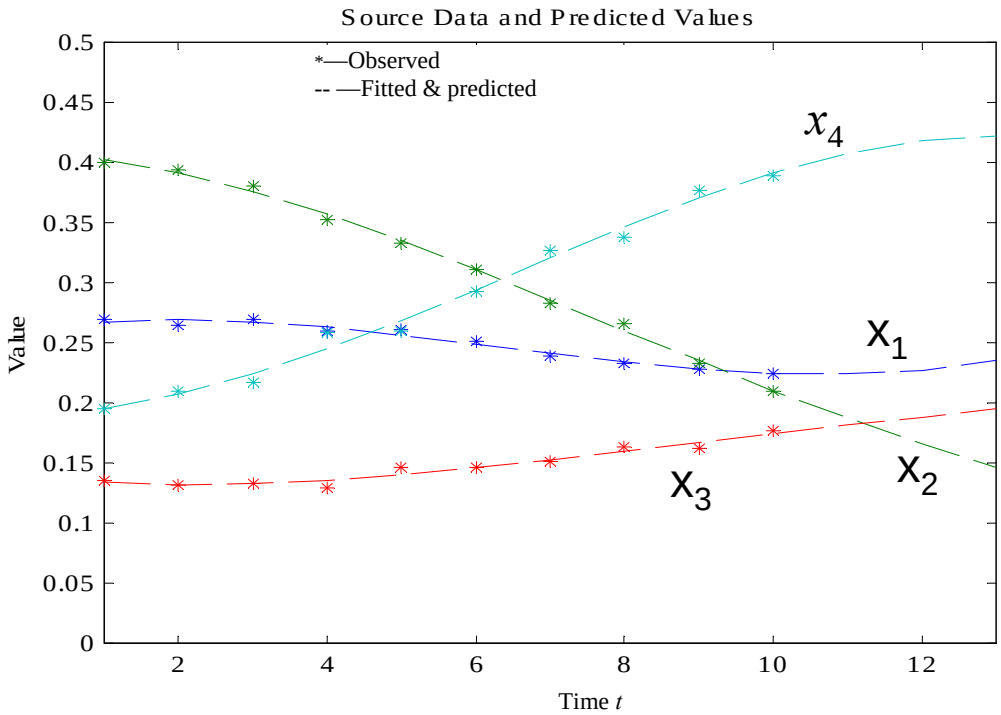


Fig. 2 Source data and predicted values in simulation

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 4 Portions of social labors in three industries of China*

Year / Industry	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
Primary industry	0.6010	0.5970	0.5850	0.5640	0.5430	0.5220	0.5050	0.4990	0.4980	0.5010
Secondary industry	0.2140	0.2140	0.2170	0.2240	0.2270	0.2300	0.2350	0.2370	0.2350	0.2300
Tertiary Industry	0.1850	0.1890	0.1980	0.2120	0.2300	0.2480	0.2600	0.2640	0.2670	0.2690

* Data source: Chinese Statistic Yearbook, 1991~2000

For t=1990,…, 1999, the computed angles θ_2^t , θ_3^t are obtained after hyperspherical mapping as listed in Table 5 and Fig.4.

Table 5 Computed angles at 1990~1999

t	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
θ_2^t	1.0328	1.0313	1.0238	1.0085	0.9968	0.9848	0.9721	0.9674	0.9689	0.9753
θ_3^t	1.1262	1.1210	1.1097	1.0923	1.0706	1.0495	1.0357	1.0312	1.0278	1.0255

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, 2001can be obtained

$$\begin{cases} \hat{\theta}_2^{2000} = 0.5065 \\ \hat{\theta}_3^{2000} = 0.9177 \end{cases} \qquad \begin{cases} \hat{\theta}_2^{2001} = 0.5017 \\ \hat{\theta}_3^{2001} = 0.9063 \end{cases}$$

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

Year / Industry	Fitting										Predicted	
	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
x_1	0.6134	0.5924	0.5733	0.5560	0.5406	0.5271	0.5156	0.5059	0.4982	0.4924	0. 4885	0. 4864
x_2	0.2096	0.2157	0.2209	0.2250	0.2283	0.2307	0.2324	0.2334	0.2337	0.2335	0. 2326	0. 2313
x_3	0.1770	0.1918	0.2059	0.2190	0.2311	0.2421	0.2520	0.2607	0.2681	0.2741	0. 2789	0. 2823
Σ	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1. 0000	1. 0000

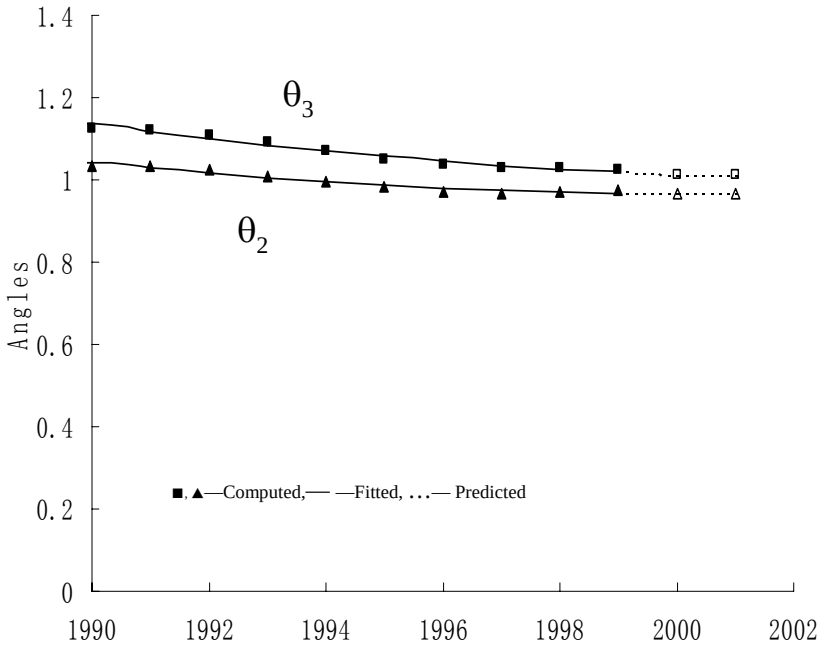


Fig. 4 Computed angles and predicted angles

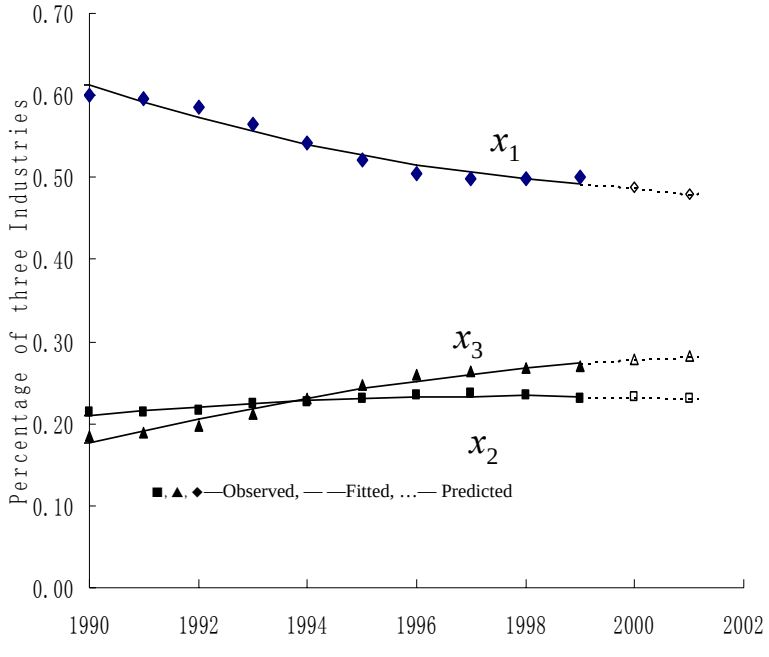


Fig. 5 Observed and predicted data of Three Industries

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;

$$Y = \left(y_1, y_2, \cdots, y_q\right)' \qquad \sum_{j=1}^{q} y_j = 1 \ , \qquad 0 < y_j < 1$$

$$X = \left(x_1, x_2, \cdots, x_p\right)' \qquad \sum_{j=1}^{p} x_j = 1 \ , \qquad 0 < x_j < 1$$

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**

$$u_j = \log \frac{y_j}{\sqrt[q]{\prod_{i=1}^{q} y_j}} \quad , j = 1,2,\cdots,q$$

$$v_j = \log \frac{x_j}{\sqrt[p]{\prod_{i=1}^{p} x_j}} \quad , j = 1,2,\cdots,p$$

Problem on Regression:

The independent variables are completely correlated.

Step 2. **Construct the PLS regression model.**

$$\hat{u}_j = \alpha_{j0} + \alpha_{j1} v_1 + \cdots + \alpha_{jp} v_p , \quad j = 1,2,\cdots,q$$

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

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

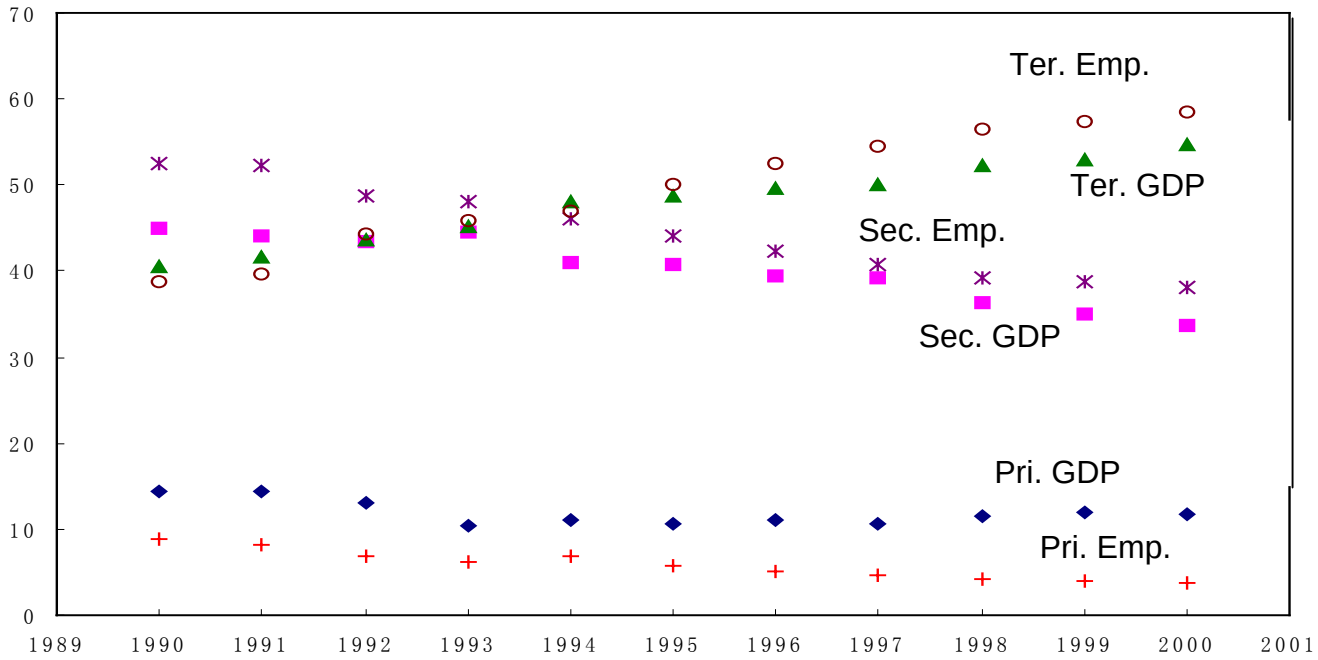
$$\hat{w}_j = \hat{u}_j - \hat{u}_q , \quad j = 1,2,\cdots,q-1$$

$$\hat{y}_q = \left\{1/(1+\sum_{i=1}^{q-1} e^{\hat{w}_i})\right\}$$

$$\hat{y}_j = e^{\hat{w}_j}\hat{y}_q, \quad j = 1,2,\cdots,q-1$$

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.

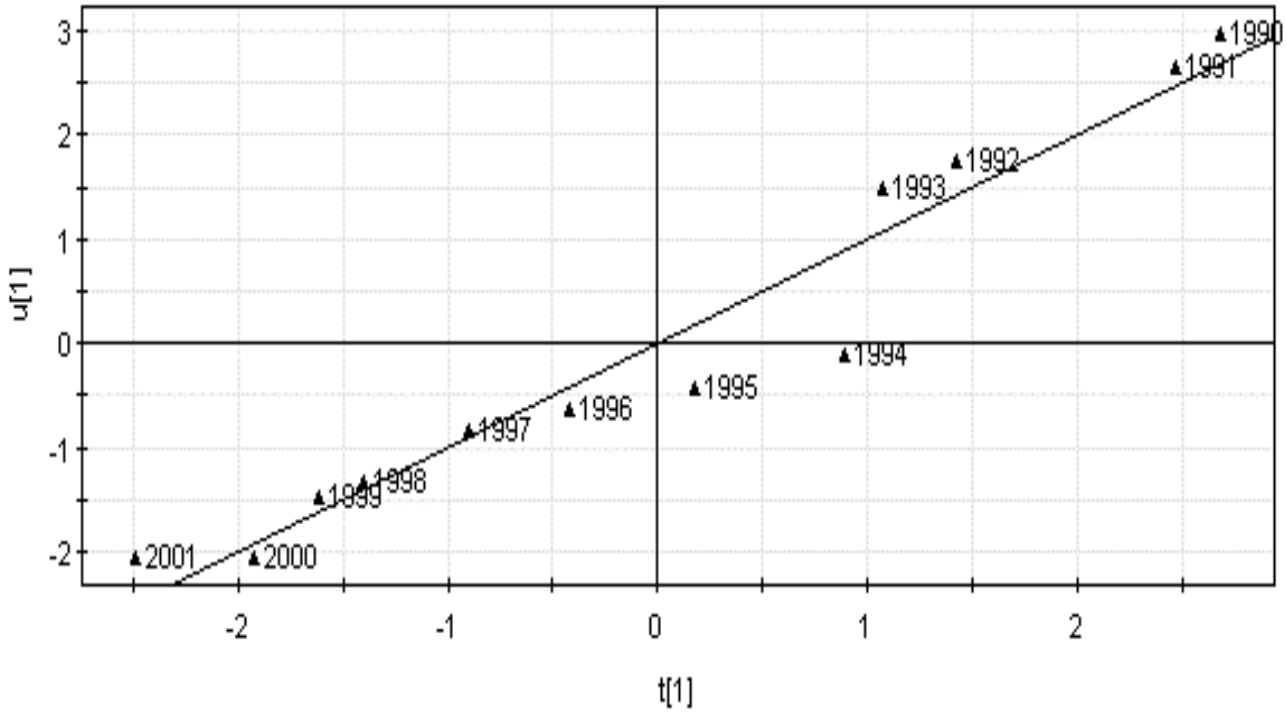


Employment portion and GDP portion in Beijing

	Employment portion (%)			GDP portion (%)		
Year	Y_1	Y_2	Y_3	X_1	X_2	X_3
1990	14.46	44.91	40.63	8.76	52.39	38.85
1991	14.32	44.12	41.56	8.15	52.18	39.67
1992	13.01	43.37	43.62	6.87	48.78	44.35
1993	12.29	43.55	44.16	6.21	48.03	45.76
1994	11.02	40.98	48.01	6.81	46.15	47.04
1995	10.61	40.73	48.65	5.84	44.1	50.06
1996	10.98	39.40	49.62	5.17	42.28	52.55
1997	10.69	39.28	50.03	4.69	40.8	54.51
1998	11.49	36.32	52.19	4.3	39.12	56.58
1999	12.04	34.95	53.01	4	38.7	57.3
2000	11.77	33.62	54.61	3.7	38	58.5
2001	11.31	34.33	54.36	3.3	36.2	60.5

PLS Regression Results

There is a strong correlation relationship between the 1_{th} PLS components t_1 and u_1.

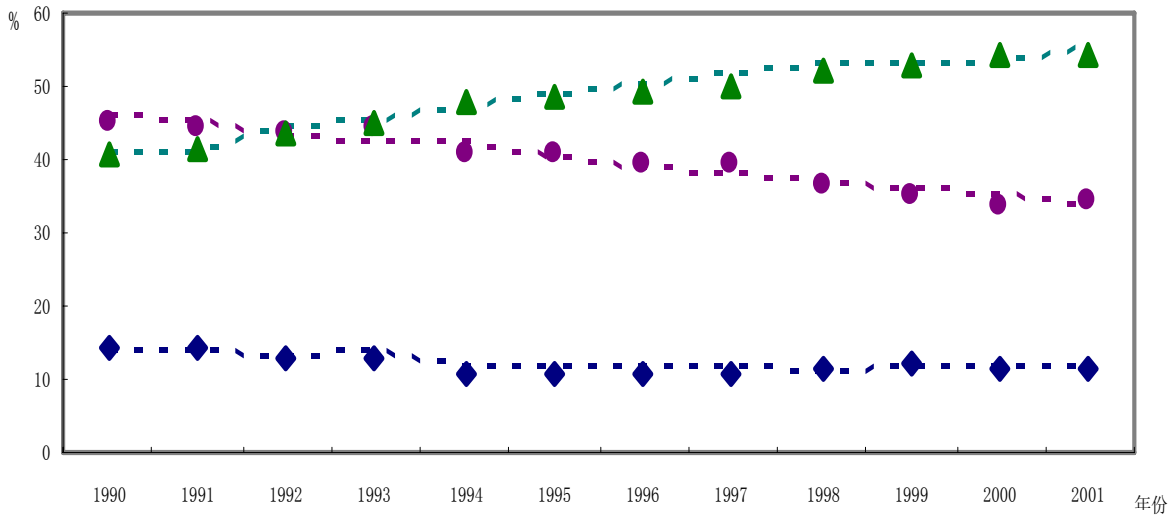


With PLSR, according to cross-validation, we select two component from X, which can explain 81.4% of explanatory variables and 88% of response, with the following models. And the Goodness of Fit of these three models are 81.3%, 88.5% and 97.0% respectively.

$$Y_1=16.5975-2.86001X_1+1.51719X_2-2.00526X_3$$

$$Y_2=28.448+0.725227X_1+0.14436X_2-0.0768685X_3$$

$$Y_3=28.1363+0.460484X_1-0.658354X_2+0.780977X_3$$



V. Multiple Linear Regression on Compositional Data

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

Y Employment Proportion			X_1 GDP Proportion			X_2 Investment Proportion		
emp1	emp2	emp3	GDP1	GDP2	GDP3	inv1	inv2	inv3

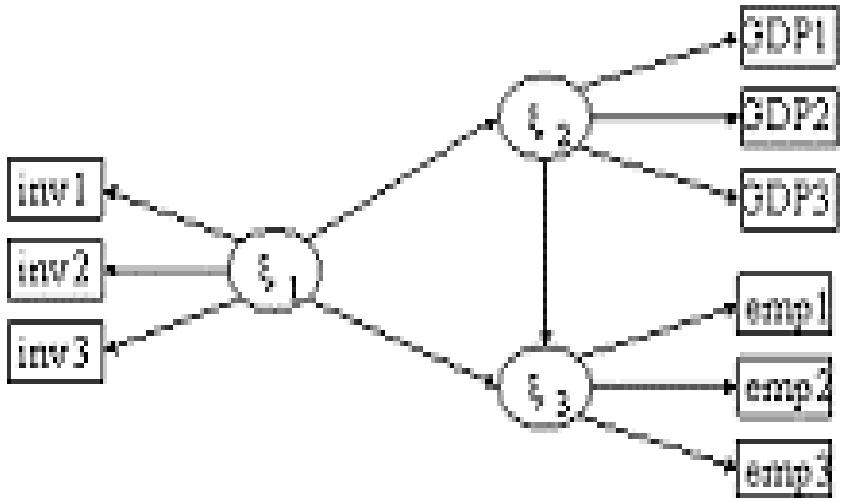
$$\hat{Y} = f(x_{11}, x_{12}, x_{13}, x_{21}, x_{22}, x_{23})$$

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

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)

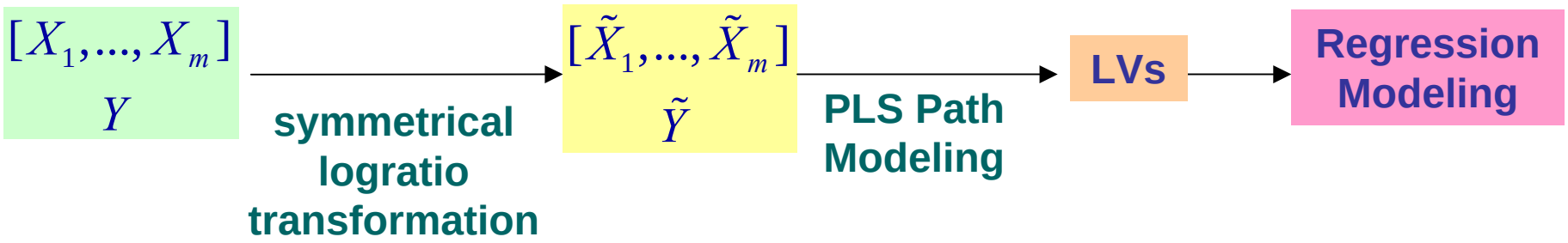
To solve the problem of hierarchy relationship of independent variables, PLS Path modeling method is applied in MCDM.

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.



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.

Table1. Definitions of LVs and MVs

LVs	Investment Proportion (inv., ξ_1)	GDP Proportion (GDP, ξ_2)	Employment Proportion (emp., ξ_3)
MVs	inv_1	GDP_1	emp_1
	inv_2	GDP_2	emp_2
	inv_3	GDP_3	emp_3

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)

	inv.	GDP	emp.
inv.	0	0	0
GDP	1	0	0
emp.	1	1	0

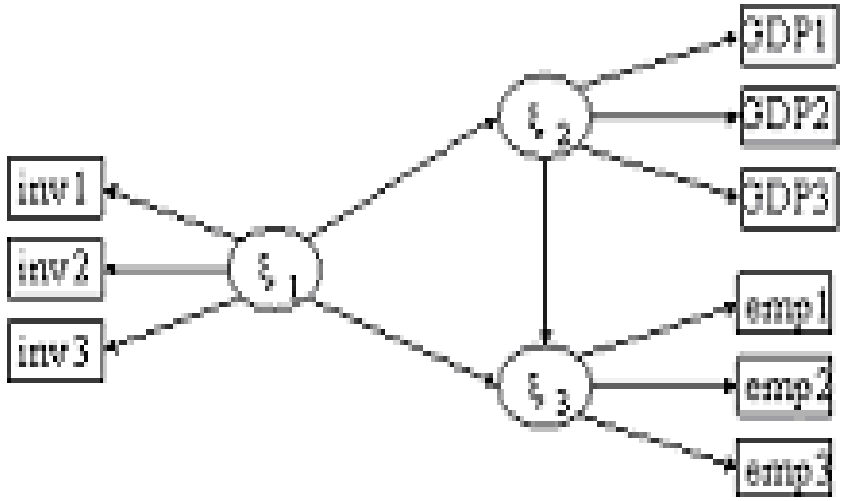


Fig.1. Path model of investment, GDP and employment

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

Table3. Path coefficients of LVs (1)

	inv.	GDP	emp.
inv.	0	0	0
GDP	0.872	0	0
emp.	0.004	0.948	0

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.

Table4. Inner design matrix(2)

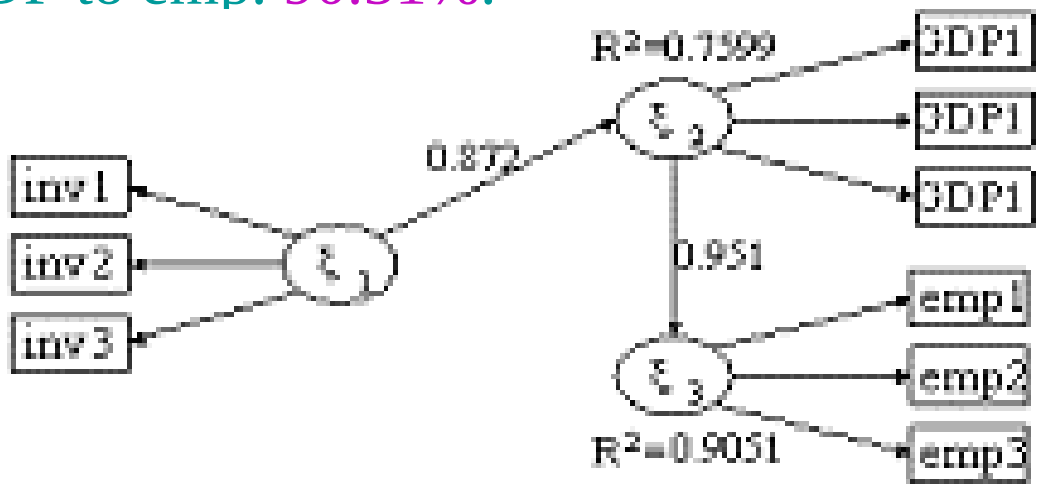
	inv.	GDP	emp.
inv.	0	0	0
GDP	1	0	0
emp.	0	1	0

Table5. Path coefficients of LVs (2)

	inv.	GDP	emp.
inv.	0	0	0
GDP	0.872	0	0
emp.	0	0.951	0

The indirect effect of inv. to emp. through GDP is **0.872×0.951=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

1. Introduction

- 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.

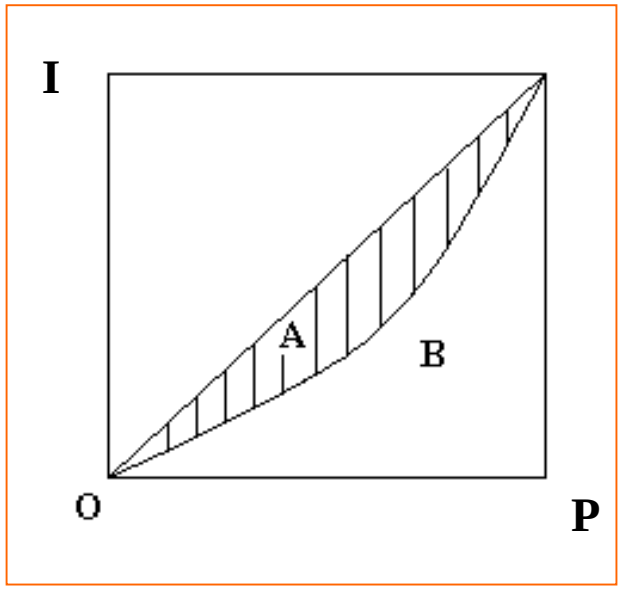


Fig. 1 Model geometry of Lorenz Curve

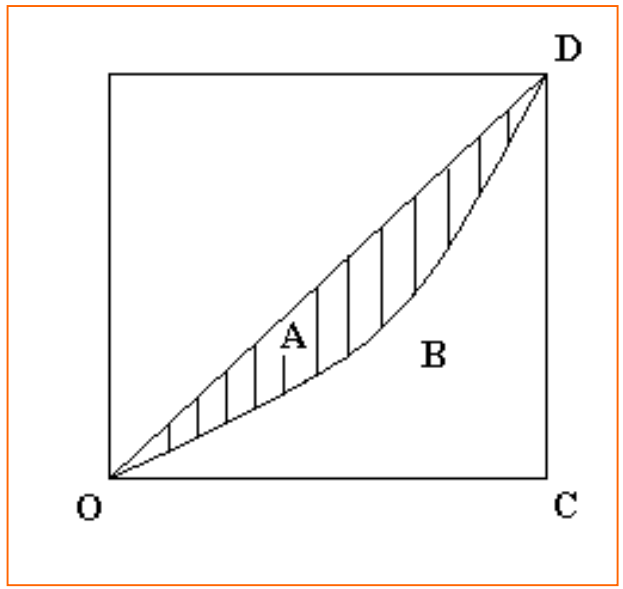
Gini Coefficient

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

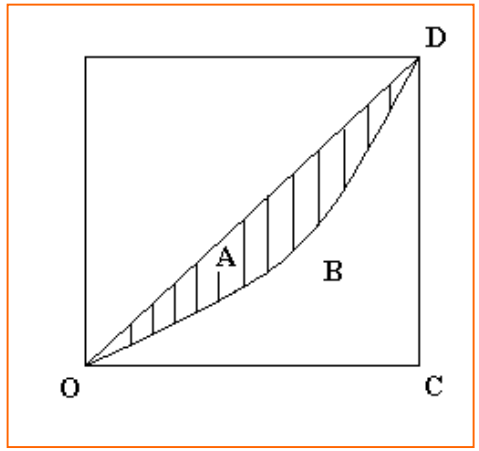
Step1. By use of predictive modeling of compositional data , we predict the future population distribution and income distribution.

Step 2. Regress a 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

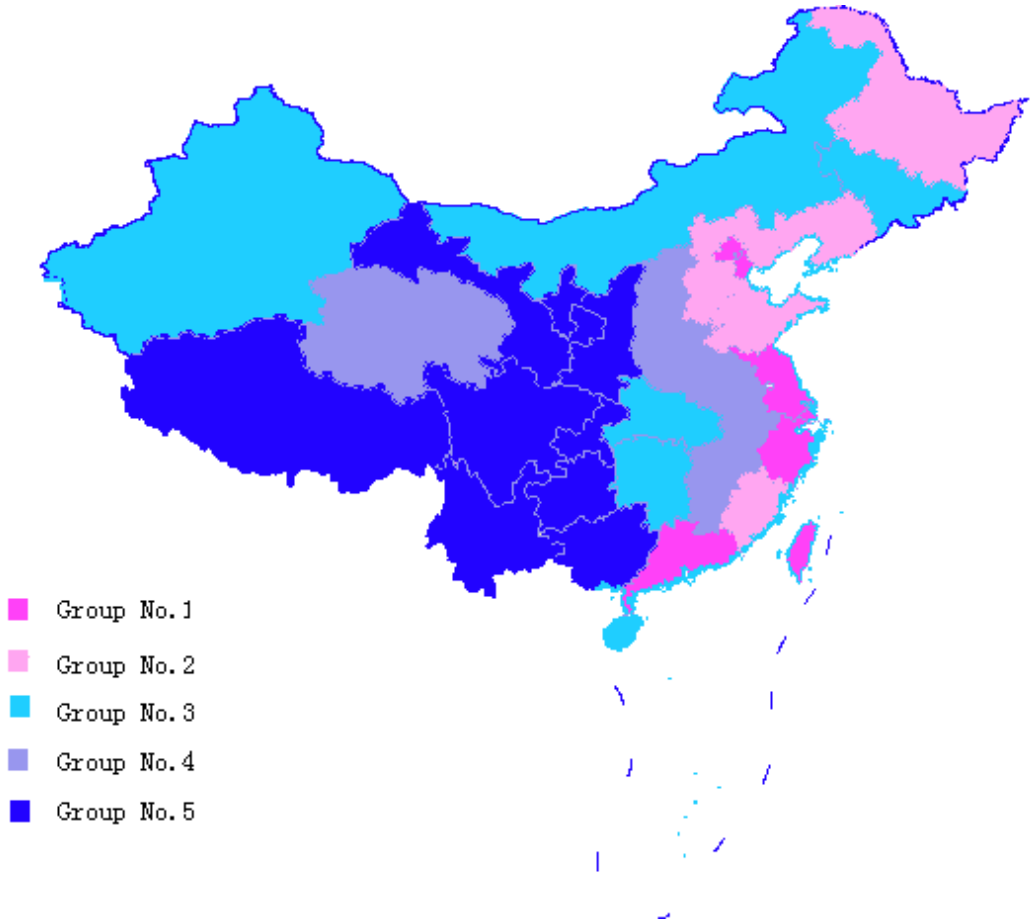
The imbalance of economic development in the different regions is one of the remarkable issues in China.

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.

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

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.

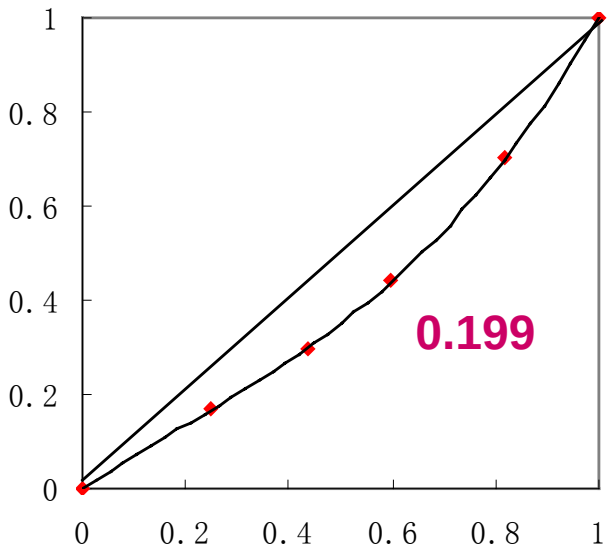


Fig. 2 Chinese population-GDP Lorentz Curve in 1991

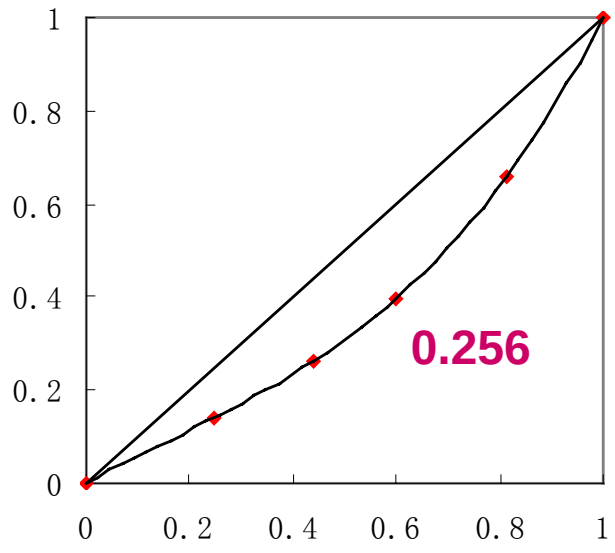


Fig. 3 Chinese population-GDP Lorentz Curve in 2001

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.

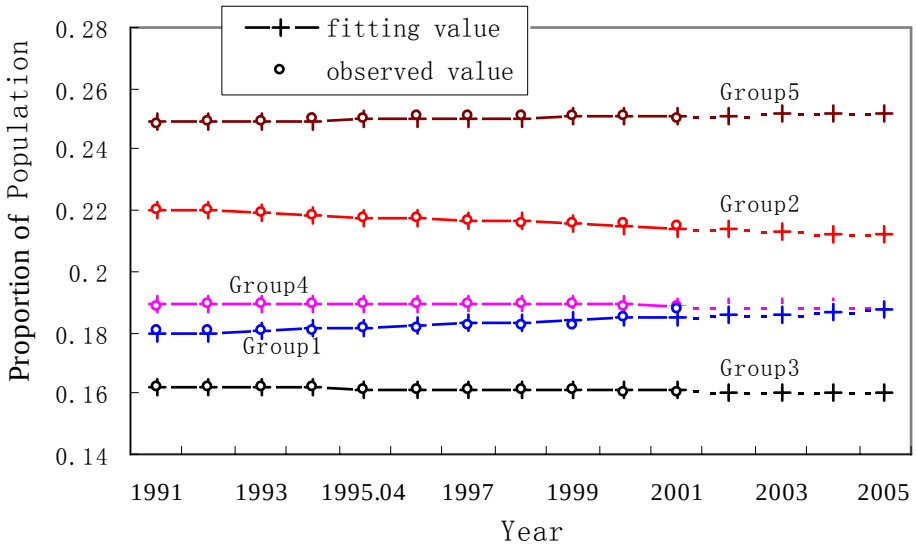


Fig.3 Fitting values of population proportion of five regions

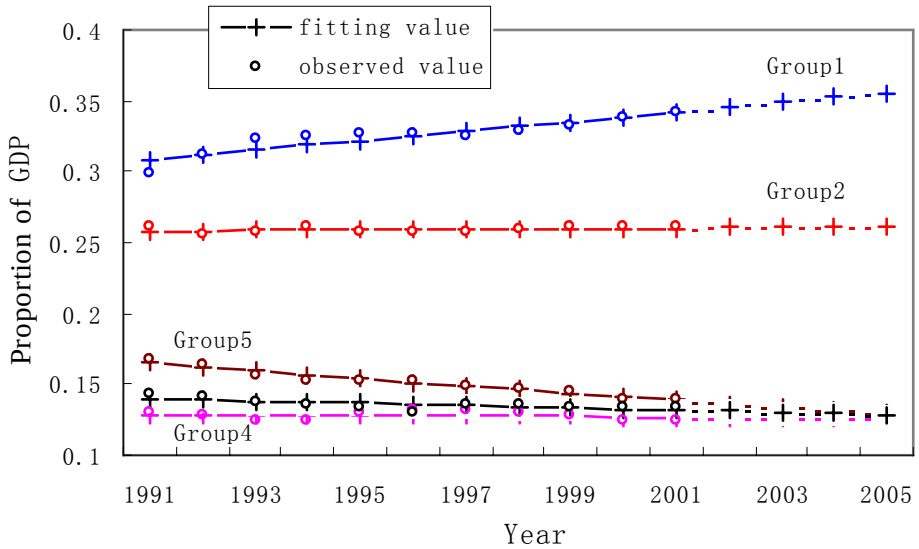


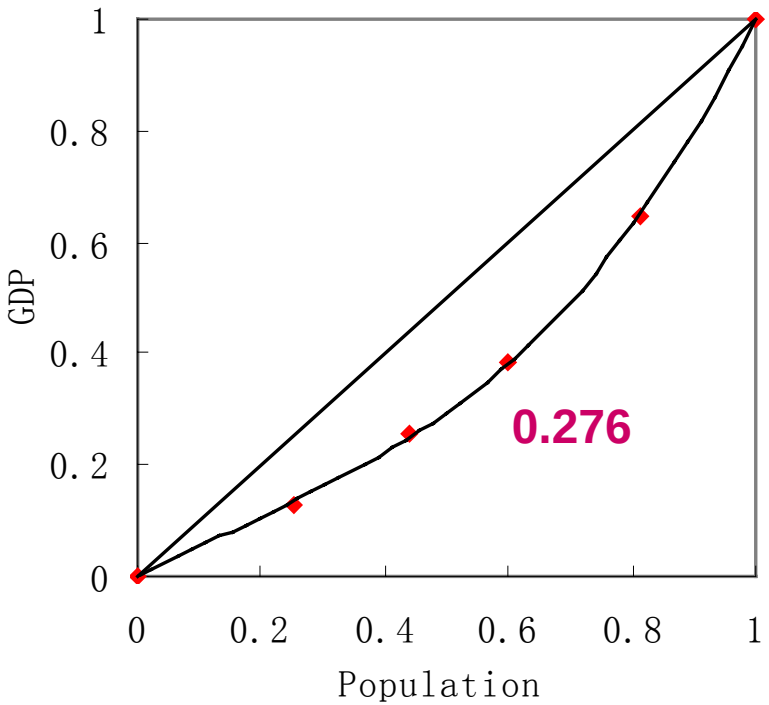
Fig. 4 Fitting values of proportion GDP of five regions

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.

Tab2. Predictive proportions of population and GDP in 2005

	Group No.1	Group No.2	Group No.3	Group No.4	Group No.5
Population rate	0.1874	0.2121	0.1602	0.1886	0.2518
GDP rate	0.3557	0.2606	0.1287	0.1257	0.1293



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.

Reference

1. Aitchison, J. . The Statistical Analysis of Compositional Data, Chapman and Hall, London.1986.
2. Wang H, Meng J. and Tenenhaus M., Regression Modeling Analysis on Compositional Data. Preceding of the 4th International Symposium on PLS and Related Methods, Spain, 2005.9.
3. Wang H., Huang W., Liu Q., Predictive Modeling of Industrial Structure, Preceding of the 7th International Conference on Industrial Management. Japan,2004.11.
4. Wang H.,Liu Q., Mok H.M.K. and Fu L. , A Hypershaerical – Transformation Forecasting Model with Compositional Data, European Journal of Operation Research, Accepted.
5. Wang H., Long W, , Liu Q., Predictive Modeling of Gini Coefficient of GDP in China, Preceding of the 5th International Conference on Management, Macao, 2004.5.