Communications in Statistics - Simulation and Computation

Latent Regression with Constrained Parameters to Determine the Weight Coefficients in Summary Index Model

Qiaoling Tong\textsuperscript{a}, Xuecheng Zou\textsuperscript{a}, Qian Ding\textsuperscript{b}, Tianzhen Liu\textsuperscript{b} & Hengqing Tong\textsuperscript{c}

\textsuperscript{a} Department of Electronic Science & Technology, Huazhong University of Science & Technology, Wuhan, China
\textsuperscript{b} School of Urban Design, Wuhan University, Wuhan, Hubei, China
\textsuperscript{c} Department of Mathematics, Wuhan University of Technology, Wuhan, China

To cite this article: Qiaoling Tong, Xuecheng Zou, Qian Ding, Tianzhen Liu & Hengqing Tong (2013): Latent Regression with Constrained Parameters to Determine the Weight Coefficients in Summary Index Model, Communications in Statistics - Simulation and Computation, 42:7, 1628-1642

To link to this article: http://dx.doi.org/10.1080/03610918.2012.671876

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: http://www.tandfonline.com/page/terms-and-conditions

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.
Latent Regression with Constrained Parameters to Determine the Weight Coefficients in Summary Index Model

QIAOLING TONG,1 XUECHENG ZOU,1 QIAN DING,2 TIANZHEN LIU,2 AND HENGQING TONG3

1Department of Electronic Science & Technology, Huazhong University of Science & Technology, Wuhan, China
2School of Urban Design, Wuhan University, Wuhan, Hubei, China
3Department of Mathematics, Wuhan University of Technology, Wuhan, China

The constructions and algorithms of summary index models to determine unknown weight coefficients using latent variable regression with constrained parameters are presented. The summary index model with a single category can be seen as a structural equations model with a latent variable without latent equations. This article gives a new suitable algorithm for it based on factor analysis, path analysis, and prescription constraint. Another summary index model is one in which all samples have been divided into multiple categories. This article also gives a suitable algorithm by alternating projection between two convex sets and prescription constraint.

Keywords Generalized linear model; Latent variable regression; Prescription constraint; Summary index model; Weight coefficient

Mathematics Subject Classification 62J05; 62J12; 62P25

1. Introduction

Many social or economic indexes, such as the Customer Satisfaction Index (CSI) (Mateos-Aparicio, 2011; Wang and Tong, 2007) and the Consumer Price Index (CPI) (Li and Hu, 2011), are summarized by some component indexes. Many mental scales, such as the Interaction Anxiety Scale (IAS) (Gore et al., 2002) and the Self-rating Anxiety Scale (SAS) (Olatunji et al., 2006), need to be summarized. A summary index model summarizes some component indexes as a main index with unknown weight coefficients that will be calculated by sample. Suppose there are \( p \) component indexes \( x_1, \ldots, x_p \) in a statistical table or a mental scale. We need to get a main index \( \xi \):

\[
\xi = \omega_1 x_1 + \cdots + \omega_p x_p
\]

where \( \omega_1, \ldots, \omega_p \) are weight coefficients. In many cases, these weight coefficients are given beforehand by an authoritative department. If the weight coefficients are unknown and calculated by sample, we call the model a summary index model.

Received December 11, 2011; Accepted February 28, 2012
Address correspondence to Dr. Tianzhen Liu and Dr. Qian Ding, School of Urban Design, Wuhan University, Wuhan 430070, Hubei, China; E-mail: bottraltz@126.com, qianding2011@yeah.net
Table 1 presents a summary of price indexes, with data coming from the BRICS Joint Statistical Publication 2011 edited by the National Bureau of Statistics of China (http://www.stats.gov.cn/english/statisticaldata/otherdata/brics2011/P020110412518703249470.pdf).

Table 1 form a $5 \times 9$ matrix. Now, we want to know the average percentage for each country from 2002 to 2010. A simple method is to take the arithmetic average, that is, $(\omega_1 = \cdots = \omega_9) = 1/9$ in (1). It is not reasonably certain, however, even though the sum of all weight coefficients equals 1. Using this method we get $\xi' = (6.5333, 2.3222, 6.5111, 11.60, 6.0444)$. If we calculate the weight coefficients using a regression method by sample, we would have our summary index model, which is our objective. We give the computation results for Table 1 at the end of the next section.

If we transpose the rows and columns in Table 1, we get a $9 \times 5$ matrix. Now, we want to know the average percentage for each year. Similar to what we had done earlier, we can take the arithmetic average, that is, $\omega_1 = \cdots = \omega_5 = 1/5$ in (1). In this way, we get $\xi' = (8.140, 6.80, 5.480, 5.580, 4.80, 6.40, 9.160, 6.60, 6.460)$. Again, this is not reasonably certain.

There are many methods to determine the weight coefficients in a summary index model. These methods usually include subjective methods, such as the equant method and the expert designated (Delphi) method, and objective methods, such as the fuzzy mathematical method (Torabi et al., 2007), the grey system method (Tien, 2009), the variation coefficient method (Wilson, 2002), the analytic hierarchy process (Saaty, 2008), and so on. Objective methods are not only more scientific but are also used as references of subjective determining methods. Our method using latent variable regression with prescription condition is a new method for summary index models.

A summary index model belongs to a kind of latent variable model (Christensen et al., 2004), and its algorithm belongs to generalized least squares (GLS) (Ho and Singer, 2001; Lidush and Neil, 2003), because the dependent variable $\xi$ is unknown. The methods of latent regression in this article are constraint regression methods (Cano and Salmerón, 2007). The equations of latent variable regression are indeterminate equations, and some suitable constraint conditions may be needed. In fact, reasonable constraint conditions are valuable and skillful.

Prescription constraint (each coefficient is nonnegative and their sum is 1) and modular constraint are the main tools in our methods. Fang and He (1985) and Fang and colleagues (1982) offered the algorithm for prescription regression. The evaluation model (Tong, 1993) is a GLS regression model with prescription condition. Wang and Tong (2007) used these ideas in SEM, and found the best iterative initial values of the partial least squares (PLS) algorithm. Tong et al. (2010) proposed a definite linear algorithm for SEM. This article
carries over the above studies and solves the problem successfully by using SEM and the GLS regression model, recurring to the ideas of path analysis and factor analysis, combining them with modular constraint, prescription constraint, and the least squares method. In this way, we offer a new algorithm to determine weight coefficients in mental scales or in summary index models.

This article discusses two kinds of summary index models. One has all samples belonging to a unique category. In this case, Eq. (1) can be seen as an SEM containing a unique structural variable, but the PLS method is not suitable to it, because there are no latent equations. This article gives a new suitable algorithm by prescription constraint and modular constraint. The other summary index model has all samples divided into multiple categories. In this case, the weight coefficients $\omega_1, \ldots, \omega_p$ in Eq. (1) must be suitable to all categories. Obviously this is more difficult. This article successfully gives a new suitable algorithm too, using prescription regression and alternating projection between two convex sets.

2. Summary Index Model with a Single Category and Constrained Solution

Assume that there are $p$ indexes $x_i, i = 1, \ldots, p$ in a scale, and each index has $N$ observations, so we have an $N \times p$ data matrix. We hope to get the summary index $\xi$, which is latent, and an $N$ dimensional vector. The weight coefficients to be estimated for every variable are $\omega_i, i = 1, \ldots, p$. Therefore, the equation is

$$\xi = \sum_{i=1}^{p} \omega_i x_i + \varepsilon$$

where $\varepsilon$ is an $N$ dimensional random error vector, $\xi$ is unknown, and $\omega_i, i = 1, \ldots, p$ is also unknown. We notice that Eq. (2) is an indefinite equation, so, it has not been applied to the summary of weight coefficients so far. We need to give a reasonable constraint, so as to give a definite algorithm, and thus a definite solution can be obtained. The data structure of this problem can be expressed as in Table 2.

First of all, we build an SEM containing a unique structural variable. We should be aware, however, that the PLS method is not suitable to this case, because there are no latent equations. Based on the idea of path analysis, we think that each observation variable not only has an impact on the latent variable as in Eq. (2), but is also impacted by the latent

<table>
<thead>
<tr>
<th>Table 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data structure of model with single category</td>
</tr>
<tr>
<td>Index</td>
</tr>
<tr>
<td>Weight coefficient</td>
</tr>
<tr>
<td>Data row</td>
</tr>
<tr>
<td>Data row</td>
</tr>
<tr>
<td>:</td>
</tr>
<tr>
<td>Data row</td>
</tr>
</tbody>
</table>
variable, as shown in the following equation:

\[
\begin{pmatrix}
  x_1' \\
  \vdots \\
  x_p'
\end{pmatrix}
= 
\begin{pmatrix}
  \lambda_1 \\
  \vdots \\
  \lambda_p
\end{pmatrix}
\xi' 
+ 
\begin{pmatrix}
  \zeta_1' \\
  \vdots \\
  \zeta_p'
\end{pmatrix}
\]  \\
(3)

where \( \zeta_i, i = 1, \ldots, p \) are \( N \) dimensional random error vectors, latent variable \( \xi \) is unknown, and load coefficients \( \lambda_i, i = 1, \ldots, p \) are also unknown. So, (3) is also an indefinite equation. Notice the dimensions of Eq. (3). On its left side, each row \( x_i' \) is a vector with \( N \) dimensions, and on its right side, each row in the first item is a number \( \lambda_i \) that multiplies a row vector \( \xi' \).

A latent variable needs to meet a number of linear relationships with observation variables, which is contradictory. In this case, we should use the least squares method and add some constraint conditions. Let \( X = (x_1, x_2, \ldots, x_p) \), \( \lambda' = (\lambda_1, \lambda_2, \ldots, \lambda_p) \), and \( \zeta = (\zeta_1, \zeta_2, \ldots, \zeta_p) \). Then, \( X \) and \( \zeta \) are both \( N \times p \) matrixes, and Eq. (3) becomes

\[
X' = \lambda \xi' + \zeta'
\]  \\
(4)

If we assume that \( \xi \) and \( \zeta \) are independent, then \( E(X'X) = E(\lambda \xi' \xi \lambda' + E(\zeta' \zeta)) \). Furthermore, if we assume the latent variable \( \xi \) is a unit vector, that is, \( \xi' \xi = 1 \), we have

\[
X'X \approx \lambda \lambda' + E(\zeta' \zeta)
\]  \\
(5)

and can obtain the modular constraint least squares solution (MCLS).

In general, we can assume that \( E(\zeta' \zeta) \) is a diagonal matrix with \( E(\zeta' \zeta) = \text{diag}(\varphi_1^2, \varphi_2^2, \ldots, \varphi_p^2) \). Both \( p \times p \) matrixes are almost equal under the meaning of the least squares, which can be expressed as follows:

\[
\begin{pmatrix}
  x_1'x_1 & \cdots & x_1'x_p \\
  \vdots & \ddots & \vdots \\
  x_p'x_1 & \cdots & x_p'x_p
\end{pmatrix}
\approx 
\begin{pmatrix}
  \lambda_1^2 + \varphi_1^2 & \cdots & \lambda_1 \lambda_p \\
  \vdots & \ddots & \vdots \\
  \lambda_p \lambda_1 & \cdots & \lambda_p^2 + \varphi_p^2
\end{pmatrix}
\]  \\
(6)

Note that the elements in the left matrix are the product of two vectors while the elements in the right matrix are the product of two numbers. Choosing the equal elements in the diagonal, we can get:

\[
x_i'x_i \approx \lambda_i^2 + \varphi_i^2, \quad i = 1, \ldots, p,
\]  \\
(7)

According to the method of factor analysis, we let the matrix \( \Sigma_x = E(X'X) \) and \( \hat{\Sigma}_x = X'X \). Suppose the diagonal elements of \( \hat{\Sigma}_x^{-1} \) are \( \hat{\phi}_i, i = 1, \ldots, p \). We can take the estimations

\[
\hat{\phi}_i^2 = \hat{\phi}_i^{-1}, \quad i = 1, \ldots, p
\]  \\
(8)

where \( \{\hat{\phi}_i\} \) is the diagonal element of the matrix \( (X'X)^{-1} \). As soon as we get \( \{\hat{\phi}_i^2\} \), we can easily estimate \( \{\lambda_i^2\} \) from (7):

\[
\hat{\lambda}_i^2 = x_i'x_i - \hat{\phi}_i^2, \quad i = 1, \ldots, p
\]  \\
(9)

Now we continue to estimate latent variable \( \xi \) by estimating its elements one by one. The estimate of \( \lambda \) from (9) contains the affection of \( \zeta \), so (4) may be expressed as an
approximate equation $X' \approx \lambda \xi'$. Using a vector $\lambda' = (\lambda_1, \ldots, \lambda_p)$ to multiply both sides of $X' \approx \lambda \xi'$, we can get

$$A \xi_j = \lambda_1 x_{1j} + \cdots + \lambda_p x_{pj}, \quad j = 1, \ldots, N$$

(10)

where $A = \lambda_1^2 + \cdots + \lambda_p^2$. So, we can get the estimation of latent variable $\xi$:

$$\hat{\xi}_j = \frac{\hat{\lambda}_1}{A} x_{1j} + \cdots + \frac{\hat{\lambda}_p}{A} x_{pj}, \quad j = 1, \ldots, N$$

(11)

Since we have the estimations of the latent variable $\xi$, we can return to (2) and make use of ordinary multivariate linear regression to estimate weight coefficients $\omega_i, i = 1, \ldots, p$.

We have to explain the reasonableness of the initial MCLS. Equation (3) is an indefinite equation, and the solution is not unique. If $\lambda_i, \xi$ is a group of solutions, $c\lambda_i, c\xi$ is also a solution, where $c$ is any nonzero constant. Therefore, we can get the solution under the condition that $\xi$ is a unit vector. The direction of any solution vectors meeting with the least squares is the same as the MCLS vector, with only their modular lengths being different.

However, which kind of modular length is appropriate? We need to add new constraints in which the weight coefficients satisfy the prescription condition, that is,

$$\sum_{i=1}^{p} \omega_i = 1, \quad \omega_i \geq 0, \quad i = 1, \ldots, p$$

(12)

In the case the equations are indeterminate equations, we have a simple method to satisfy prescription condition.

If we do not consider random errors in (4), (5), (6), and (9), the algorithm in this section is simple and easy to program. We have an $N \times p$ observation data matrix $X$. We take an array $\Lambda = (\lambda_1^2, \ldots, \lambda_p^2) = diag(X'X)$. Let $A = \lambda_1^2 + \cdots + \lambda_p^2$, and an array $B = (\frac{\lambda_1}{A}, \ldots, \frac{\lambda_p}{A})$. Let $b = \frac{\lambda_1}{A} + \cdots + \frac{\lambda_p}{A}$. Normalizing $B$, we get an array $\Omega = B/b = (\omega_1, \ldots, \omega_p)$. Those are the weight coefficients we needed.

We use this method to calculate the data in Table 1, and get the following results: $(\omega_1, \ldots, \omega_9) = (0.1355, 0.1206, 0.1028, 0.1076, 0.0959, 0.1020, 0.1250, 0.1122, 0.0984)$, $\xi' = (6.7667, 2.2481, 6.4430, 11.8728, 6.2750)$. These results are somewhat different from the arithmetic average.

We use this method to calculate the data in the transpose of Table 1, and get the following results: $(\omega_1, \ldots, \omega_5) = (0.2262, 0.1593, 0.1776, 0.2258, 0.2113)$, $\xi' = (8.9170, 6.0175, 5.6645, 4.3485, 6.0407, 6.6353, 10.6667, 5.4985, 6.7403)$. These results are also somewhat different from the arithmetic average.

3. Summary Index Model with Multiple Categories and Constrained Solution

Some samples in the summary index model can be divided into several categories. For example, the objects in an anxiety scale can be classified in accordance with the degree of anxiety. We want to summarize the anxiety degree of each type of sample into one score, not two or more scores, but we cannot determine the degree of anxiety in advance. How do we get the weight coefficients now? We should build a new model.

In our model, we assume that there are $p$ indexes and $m$ categories. For each index of each category, there are $n$ mental records. All samples should have $N = mn$ observation
records, forming an $N \times p$ matrix. In our model, the weight coefficients are to be determined and have not been set in advance.

The $p$ indexes are variables expressed by $x_{i(1)}, \ldots, x_{i(p)}$, respectively. The element $x_{ijk}$ denotes the mental records given by the $i^{th}$ ($i = 1, \ldots, n$) respondent on the $j^{th}$ ($j = 1, \ldots, p$) index of the $k^{th}$ ($k = 1, \ldots, m$) object. Thus, we obtain a matrix $X_{(mn \times p)} = \{x_{ijk}\}$. For the $k^{th}$ object, we must give one and only one terminal score $y_k$ ($k = 1, \ldots, m$). These scores are unknown before solving. The model is different from the ordinary regression model if we build a regression model and take $y_k$ ($k = 1, \ldots, m$) as the dependent variables. Table 3 shows the data structure of the model.

In this data structure, the dependent variable is unknown, so the model is a generalized linear model. To use a regression model to obtain the estimations of $\beta_1, \ldots, \beta_p$ and $y_1, \ldots, y_m$, it is necessary to give a constraint on weight coefficients $\beta_j$ ($j = 1, \ldots, p$). Obviously, these weight coefficients should satisfy $\beta_j \geq 0$, ($j = 1, \ldots, p$) and $1_p' \beta = 1$, where $1_p = (1, \ldots, 1)'$, $\beta = (\beta_1, \ldots, \beta_p)'$, that is, $\beta_1 + \cdots + \beta_p = 1$. This is a prescription constraint (Fang et al., 1982; Fang and He, 1985). Another constraint on the dependent variable has been elaborated above, that is, we can give only one score to each group. We thus define the dependent variable $Y = Dy$, where $Dy = y \otimes 1_n$, the definition of $D$ is $D_{mn \times m} = I_m \otimes 1_n$, $y = (y_1, \ldots, y_m)'$, and $\otimes$ is the Kronecker product.

The model can then be expressed as follows (Tong, 1993):

$$Dy = X\beta + \varepsilon, E(\varepsilon) = 0, Var(\varepsilon) = \sigma^2I \quad (13)$$

$$1_p' \beta = 1, \beta \geq 0, D_{mn \times m} = I_m \otimes 1_n \quad (14)$$

Now we discuss the least squares estimation (LSE) of the model in three steps.

Firstly, if $Dy$ is known, it is just an ordinary constraint regression model. Only considering the constraint $1_p' \beta = 1$, we can get an explicit solution by the method of Lagrange

| Table 3 |
|-------------------|--------|--------|--------|--------|
| Index      | Weight coefficient | $x_{i(1)}$ | $x_{i(2)}$ | $\ldots$ | $x_{i(p)}$ |
| Category 1 | $y_1$ | $x_{111}$ | $x_{121}$ | $\ldots$ | $x_{1p1}$ |
| Category 2 | $y_1$ | $x_{211}$ | $x_{221}$ | $\ldots$ | $x_{2p1}$ |
| Category 3 | $y_1$ | $x_{n11}$ | $x_{n21}$ | $\ldots$ | $x_{np1}$ |
| Category 4 | $y_m$ | $x_{11m}$ | $x_{12m}$ | $\ldots$ | $x_{1pm}$ |
| Category 5 | $y_m$ | $x_{21m}$ | $x_{22m}$ | $\ldots$ | $x_{2pm}$ |
| Category 6 | $y_m$ | $x_{n1m}$ | $x_{n2m}$ | $\ldots$ | $x_{npm}$ |
Multiplier. Let

\[ Q(\beta, y) = (Dy - X\beta)'(Dy - X\beta) \]  \hspace{1cm} (15) \\
\[ \varphi(\beta, y) = Q - 2\lambda(1_p'\beta - 1) \]  \hspace{1cm} (16)

where \( \lambda \) is the multiplier. We have

\[ \frac{\partial \varphi}{\partial \beta} = -2nX_0'y + 2X'X\beta - 2\lambda 1_p' = 0 \]  \hspace{1cm} (17) \\
\[ \frac{\partial \varphi}{\partial y} = 2ny - 2nX_0'\beta = 0 \]  \hspace{1cm} (18)

where

\[ X_0' = \frac{1}{n}(I_m \otimes 1_n')X = \frac{1}{n}DX \]  \hspace{1cm} (19)

It is a compression matrix by taking the average value of each column for each data block in matrix \( X \). Then

\[ \hat{\beta} = X_0'\beta \]  \hspace{1cm} (20)

Let

\[ P_D = I_{mn} - \frac{1}{n}DD' \]  \hspace{1cm} (21)

It is easy to verify that \( P_D \) is a projection matrix. Since \((X'X - nX_0'X_0)\beta = \lambda 1_p'\), let

\[ A = X'X - nX_0'X_0' = X'P_DX \]  \hspace{1cm} (22)

where \( A \) is a \( p \times p \) matrix, \( A = X'P'DDX = (P DX)'(P DX) \), so \( A \) is invertible if and only if \( rk(P DX) = p \). When \( A \) is invertible, \( \beta = \lambda A^{-1}1_p' \). From \( 1_p'\beta = 1 \) we have \( \lambda 1_p' A^{-1}1_p' = 1 \). Then, the solution of \( \beta \) is

\[ \hat{\beta} = \lambda A^{-1}1_p = \frac{A^{-1}1_p}{1_p' A^{-1}1_p} \]  \hspace{1cm} (23)

Summarizing the above, we get the following theorem.

**Theorem 1.** If \( Dy \) is known and \( rk(P DX) = p \), under the constraint \( 1_p'\beta = 1 \),

\[ Q(\beta, y) = ||Dy - X\beta||^2 \xrightarrow{x,\beta} \min \]  \hspace{1cm} (24)

has unique solution (20) and (23). If each component of \( \hat{\beta} \) is nonnegative, (20) and (23) are also the solution of the model (13), (14).

Secondly, if some components of \( \hat{\beta} \) are negative, we must consider the constraints \( 1_p'\beta = 1 \) and \( \beta \geq 0 \) simultaneously. Then, it is a prescription regression model. We consider the existence and uniqueness of the solution of the model (13), (14), and express the model
as:

\[
\begin{cases}
\quad Dy = X\beta + \epsilon \\
\quad 1_p'\beta = 1 \\
\quad \beta \geq 0
\end{cases}
\] (25)

When \( y \) is known, it is the PR Model of prescription regression (Fang et al., 1982). According to the Theorem in that article (see Appendix of this article), we can conclude the following theorem.

**Theorem 2.** The existence and uniqueness of the solution of summary index model (13), (14).

1. If \( y \) is known, then the solution of summary index model (13), (14) exists;
2. If \( y \) is known, and \( rk(X) = p \), then there exists a unique solution for summary index model (13), (14);
3. If \( y \) is known, \( rk(X) = p \), and the solution of (23) \( \hat{\beta} \geq 0 \), then it is the solution of summary index model (13), (14). If some component of \( \hat{\beta} \) of (23) is negative, then the component of \( \hat{\beta} \) of summary index model (13), (14) must be on the boundary of constraint conditions, that is, it is 0.

However, the dependent variable is unknown in the summary index model, so we continue our discussion.

Thirdly, we discuss the algorithm of the model when \( Dy \) is unknown. We consider the geometric background of (24). Denote sets

\[ A = \{Dy | y \in \mathbb{R}^m\} \] (26)

\[ B = \{X\beta | 1_p'\beta = 1, \beta \geq 0, \beta \in \mathbb{R}^p\} \] (27)

where \( A \) is a linear subspace expanded by the column vectors of \( D \), and \( B \) is a convex set under the constraint \( 1_p'\beta = 1, \beta \geq 0 \) in linear subspace \( X\beta \). The formula (24) means to seek the shortest Euclidean distance between sets \( A \) and \( B \). Obviously, \( A \) and \( B \) are two closed convex sets and \( B \) is bounded. According to the theorem “the distance between two closed convex sets can be reached”, the solution of (24) exists.

How to find the shortest distance between two convex sets? We consider the method of alternating projection between two sets. Let \( A_0 \) be a point in the set \( A \), \( d(A_0, B_0) \) the Euclid distance between two points \( A_0 \) and \( B_0 \), and \( d(A_0, B) \) the shortest Euclid distance between point \( A_0 \) and convex set \( B \). If \( d(A_0, B_0) = d(A_0, B) \), we call \( B_0 \) the projection from point \( A_0 \) to set \( B \). The following is the alternating projection process.

Take an arbitrary initial value \( A_0 \in A \), and find \( B_0 \in B \), satisfying \( d(A_0, B_0) = d(A_0, B) \). For \( B_0 \), take \( A_1 \in A \), satisfying \( d(B_0, A_1) = d(B_0, A) \). For \( A_i \in A \), take \( B_i \in B \), satisfying \( d(A_i, B_i) = d(A_i, B) \). For \( B_i \), take \( A_{i+1} \in A \), satisfying \( d(B_i, A_{i+1}) = d(B_i, A) \), and so on. When \( d(A_i, B_i) < \epsilon \), the iterative process is stopped and computation is completed. The meaning of convergence of aforesaid iterative process is:

\[ \lim_{i \to \infty} d(A_i, B_i) = d(A^*, B^*) \] (28)

where \( A^* \in A \) and \( B^* \in B \), respectively. The process may be shown as Fig. 1.

For the initial value \( y_0 \) of \( y \), \( Dy_0 \) is \( A_0 \) in Fig. 1, the least squares solution of model is:

\[ Q(\beta) = ||Dy_0 - X\beta||^2 \xrightarrow{\beta} \min \] (29)
where $D_{y0}$ is a constant vector and $\beta$ has convex constraint. We can solve (29) according to Theorems 1 and 2, and we can get $\beta_0$, the estimate of $\beta$, that is $B_0$ in Fig. 1. For known $\beta_0$, the model is a common multivariable linear regression model:

$$Q(y) = ||Dy - X\beta_0||^2 \rightarrow \min$$

where $X\beta_0$ is a constant vector and $y$ has no constraint. We can get $y_1$, the solution of (30), which is $A_1$ in Fig. 1. Finally, we can get the solution of the model according to (28). The computation process shows the convergence process is very fast.

Concluding the above algorithm, we have following theorem.

**Theorem 3.** The solution of the summary index model (25) can be obtained by alternative projection algorithm between two convex sets as (29) and (30).

Now we give a data example for this section. We need to evaluate the construction qualities of three buildings. There are six indexes in the evaluation table. Eight specialists for each building are to give evaluations. Total scores form a matrix with 24 rows and 6 columns. We fold the data matrix into three parts as in Table 4.

<table>
<thead>
<tr>
<th>$x_{(1)} \sim x_{(6)}$</th>
<th>$\beta_1 \sim \beta_6$</th>
<th>$x_{(1)} \sim x_{(6)}$</th>
<th>$\beta_1 \sim \beta_6$</th>
<th>$x_{(1)} \sim x_{(6)}$</th>
<th>$\beta_1 \sim \beta_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1$</td>
<td>3 2 8 5 6 3</td>
<td>$y_2$</td>
<td>9 6 6 4 9 5</td>
<td>$y_3$</td>
<td>1 3 6 5 1 8</td>
</tr>
<tr>
<td>$y_1$</td>
<td>3 7 7 3 4 6</td>
<td>$y_2$</td>
<td>3 1 9 1 6 4</td>
<td>$y_3$</td>
<td>9 3 2 4 9 7</td>
</tr>
<tr>
<td>$y_1$</td>
<td>9 1 3 6 7 1</td>
<td>$y_2$</td>
<td>3 9 6 5 9 1</td>
<td>$y_3$</td>
<td>2 7 4 4 5 2</td>
</tr>
<tr>
<td>$y_1$</td>
<td>4 3 5 7 6 1</td>
<td>$y_2$</td>
<td>8 8 9 2 1 7</td>
<td>$y_3$</td>
<td>6 9 6 9 6 2</td>
</tr>
<tr>
<td>$y_1$</td>
<td>9 6 1 7 5 8</td>
<td>$y_2$</td>
<td>9 7 8 2 6 9</td>
<td>$y_3$</td>
<td>9 5 2 6 3 5</td>
</tr>
<tr>
<td>$y_1$</td>
<td>1 8 3 1 8 4</td>
<td>$y_2$</td>
<td>3 4 4 6 6 9</td>
<td>$y_3$</td>
<td>5 9 2 2 4 7</td>
</tr>
<tr>
<td>$y_1$</td>
<td>7 8 1 4 3 1</td>
<td>$y_2$</td>
<td>5 9 3 5 6 5</td>
<td>$y_3$</td>
<td>2 7 7 9 3 5</td>
</tr>
<tr>
<td>$y_1$</td>
<td>2 8 7 1 7 9</td>
<td>$y_2$</td>
<td>3 8 2 6 9 8</td>
<td>$y_3$</td>
<td>4 3 1 9 5 2</td>
</tr>
</tbody>
</table>
The computation can be performed by the software DASC (Data Analysis &
Statistical Computation) developed by Tong et al. (2011). DASC can be down-
loaded from the website http://public.whut.edu.cn/slx/English/index.HTM. When the file
DASC-English.rar is unzipped to a folder \DASC-English\, enter the path \DASC-
English\DASC\, and hit file DASC.exe, after which a menu screen appears. Hit the top
menu “Regression→Biased/Compress Regression→Evaluation Model with LSE”, and the
example data will be loaded into Area A on the screen. Hit the button “Calculate” in Area B,
and all calculations of the data example will be performed automatically. The computation
results can be seen in Area C of DASC.

The main computation results of this data example with Table 4 are as follows. Re-
gression coefficient \(\beta\), namely, weighted coefficients, are 0.1657, 0.1866, 0.1547, 0.1535,
0.1749, and 0.1647, respectively. Their sum is 1. Fig. 2 shows the other computation result.
The evaluation scores for each classroom are \(\hat{y}_1 = 5.1686\), \(\hat{y}_2 = 5.1263\), and \(\hat{y}_3 = 5.2712\),
as three horizontal line sections in Fig. 2. The estimations of \(X\beta\) contain 24 values, omitted
here, and form the dotted line in Fig. 2.

A user can also hit top menu “Regression→Structural Equation Model(CSI, PLS)
→Summary Index Model with Single Category”, then perform the computations of Table 1
in Sec. 2.

Now, we consider the MLE of the summary index model. The main problem is that
the dependent variable is unknown. Some special construction of the dependent variable
is needed, that is, \(Dy = y \otimes n\). Changing the assumption of the random error in (13) as:

\[
Dy = X\beta + \varepsilon, \varepsilon \sim N(0, \sigma^2 I)
\]  

we have

\[
Dy \sim N(X\beta, \sigma^2 I) \tag{32}
\]

and

\[
D'^TDy \sim N(D'^TX\beta, D'^\sigma^2 D) \tag{33}
\]

Since \(D_{mn \times m} = I_m \otimes 1_n, D'D = nI_m, D'X = nX_0',\) we have

\[
y \sim N\left(X_0\beta, \frac{\sigma^2}{n}I_m\right) \tag{34}
\]  

![Figure 2. Summary index model with three categories. (Color figure available online).](image)
The density function of the independent variable $y$ is:

$$f(y) = \frac{n^\frac{m}{2}}{(2\pi)^\frac{m}{2} \sigma^m} \exp \left[-\frac{n}{2\sigma^2} (y - X_0' \beta)'(y - X_0' \beta)\right] \quad (35)$$

The logarithm likelihood function is:

$$\ln L(y, \beta, \sigma^2) = \frac{m}{2} \ln n - \frac{m}{2} \ln(2\pi \sigma^2) - \frac{n}{2\sigma^2} (y - X_0' \beta)'(y - X_0' \beta) \quad (36)$$

Since $\sigma^2$ is assumed to be known, (36) can be rewritten as:

$$\ln L(y, \beta) = c - \frac{n}{2\sigma^2} (y - X_0' \beta)'(y - X_0' \beta) \quad (37)$$

where $c = \frac{m}{2} \ln n - \frac{m}{2} \ln(2\pi \sigma^2)$. If $y$ is supposed to be known, maximizing $\ln L(y, \beta)$ is just minimizing $(y - X_0' \beta)'(y - X_0' \beta)$, and its solution is:

$$\hat{y} = X_0' \hat{\beta} \quad (38)$$

This is identical to the least squares solution (20).

On the other hand, from (32) $Dy \sim N(X\beta, \sigma^2 I_{mn})$ we have:

$$f(Dy) = \frac{1}{(2\pi)^\frac{mn}{2} \sigma^{mn}} \exp \left[-\frac{1}{2\sigma^2} (Dy - X\beta)'(Dy - X\beta)\right] \quad (39)$$

Maximizing (39) means minimizing $(Dy - X\beta)'(Dy - X\beta)$ with (38) and the prescription condition. This conclusion is the same as (24) in Theorem 1. Now we can see it as a general optimal problem, and solve it by Powell algorithm or other optimal algorithm. However, we can also solve it by alternative projection algorithm between two convex sets as in Theorem 3.

### 4. Summary and Discussion

This article considers determining the weight coefficients in a summary index model using a regression method with an unknown dependent variable and weight coefficients by sample. Two cases are examined, one where the samples come from a single category and the other where the samples come from multiple categories. Of course, it is a difficult problem. The main contribution of this article is to give new algorithms for the summary index models.

To calculate the summary index model in which samples come from a single category, we build a structural equations model with a single latent variable, and propose a definite linear algorithm using the constraint least squares method with prescription condition for regression coefficients. The prescription constraint condition is reasonable and the modular length constraint is skillful.

To calculate the summary index model in which samples come from multiple categories, we build a generalized linear regression model and use the constraint least squares method with prescription condition. We obtain the LSE and MLE of regression coefficients by alternating projection algorithm between two convex sets. The alternating projection algorithm is intuitive and effective.
Our method is an innovation in statistical model theory and computation method, which is important to improve the method of determining weight coefficients in summary index models. The models and methods are suitable not only for psychological analysis by mental scales but also for other index summarization problems such as CSI and CPI, as long as the dependent variables in the regression model are latent.

Prescription regression plays an important role in our method. The problem of prescription regression is that when a coefficient is negative, the corresponding index will be deleted. To avoid this problem, we propose, at the end of Sec. 2, a simplified algorithm ignoring random errors. We can also adopt the simplified algorithm in Sec. 3, which is based on path analysis and prescription constraint.

When we ignore random errors, the model (13) is

\[ Dy = X\beta \]  

(40)

According to the idea of path analysis, \( X \) is also affected by \( Dy_{n \times 1} \), so we have

\[ X' = \Lambda y'D' \]  

(41)

where \( \Lambda_{p \times 1} = (\lambda_1, \ldots, \lambda_p)' \). Thus, \( X'X = \Lambda y'D'Dy\Lambda' = \Lambda y'n I_m y\Lambda' = n\Lambda y'y\Lambda' \). First, we let \( y \) be a unit vector, that is, \( y'y = 1 \), and we have \( X'X = n\Lambda\Lambda' \). These are two \( p \times p \) matrices. We take the diagonal elements of two covariance matrices, and have array \( (\lambda_1^2, \ldots, \lambda_p^2) = \frac{1}{n} \text{diag}(X'X) \). Thus, we have the solution of \( \Lambda \), that is, \( \lambda_i = \frac{1}{\sqrt{n}} \sqrt{x_i} \), \( i = 1, \ldots, p \). We take left multiple for (41) with \( \Lambda' \), and get \( \Lambda'X' = \Lambda'\Lambda y'D' \). Then, we transpose it and have \( X\Lambda = aDy \), where \( a = \lambda_1^2 + \cdots + \lambda_p^2 \). Thus, we have

\[ Dy = X\frac{\Lambda}{a} \]  

(42)

Comparing this with (40) we know we have the solutions of coefficient \( \tilde{\beta} = \frac{\Lambda}{a} = (\frac{\lambda_1}{a}, \ldots, \frac{\lambda_p}{a})' \). But now \( \tilde{\beta} \) does not satisfy the prescription conditions. We notice that Eq. (40) are indefinite equations, and \( y \) need not to be a unit vector. Thus, we take the normalization of \( \tilde{\beta} = \frac{\Lambda}{a} \). Let the sum of each element of \( \tilde{\beta} \) be \( b = \frac{\lambda_1}{a} + \cdots + \frac{\lambda_p}{a} \); we get array \( \beta = \tilde{\beta}/b = (\frac{\lambda_1}{ab}, \ldots, \frac{\lambda_p}{ab})' = (\beta_1, \ldots, \beta_p)' \). These are the summary coefficients in (40); they are nonnegative and their sum is 1.

Obviously this simplified algorithm is easy to program, but its statistical properties need to be discussed further. Meanwhile, we notice that the model in Sec. 2 is merely a cell of observation equations in SEM. Our method can offer a definite linear algorithm for SEM. We should compare this algorithm with other algorithms such as PLS and Linear Structure Relationship (LISREL). Further research will be undertaken.

**Acknowledgments**

The authors are grateful to Dr. T. Krishna Kumar, whose constructive comments have led to a substantial improvement in the presentation of the article. We also thank the editor and reviewers for their insightful comments and suggestions. This study was supported by the National Natural Science Foundation of China (81271513, 51008238).
Appendix

A Theorem About the Solution of Prescription Regression Model in Fang et al. (1982)

The prescription regression model is:

\[ P R \begin{cases} Y = X\beta + \varepsilon \\ 1_p'\beta = 1, \beta \geq 0 \end{cases} \]

where \( Y \) is \( n \times 1 \), \( X \) is \( n \times p \), \( \beta \) is \( p \times 1 \), \( \varepsilon \) is \( n \times 1 \), \( p \) is \( p \times 1 \) matrix respectively, and \( 1_p = (1, \ldots, 1)' \).

Let \( b = (b_1, b_2, \ldots, b_p)' \) be the least squares estimation of \( \beta \) under the constraints \( 1_p'\beta = 1, \beta \geq 0 \), and let \( D \) be all possible solutions of \( b \), that is

\[ D = \{ b | b \geq 0, 1_p'b = 1 \} \]

It is the region of hyperplane \( 1_p'b = 1 \) in the first quadrant coordinate. When \( p = 3 \), \( D \) is shown in Fig. A1.

![Figure A1](image)

\textbf{Figure A1.} The region of all possible solutions of model PR \((p = 3)\).

\( D \) is the convex polygon in hyperplane \( 1_p'b = 1 \). For every positive integer \( r \), \( 1 \leq r \leq p \) and numbers \( 1 \leq i_1 < i_2 < \cdots < i_r \leq p \),

\[ d_{i_1,i_2,\ldots,i_r} = \{ b | \sum_{i = i_1,i_2,\ldots,i_r} b_i = 1, b_{i_1} = b_{i_2} = \cdots = b_{i_r} = 0, b_i \geq 0, i \neq i_1, i_2, \ldots, i_r \} \]

is the boundary of \( D \), and all boundaries of \( D \) are

\[ \bar{D} = \bigcup_{i = 1}^{p} d_i \]

For any \( j \neq i \), \( d_{ij} \) is the subboundary of \( d_i \), obviously

\[ d_i \supset \bigcup_{j \neq i} d_{ij} \]

For any \( k \neq i_1, i_2, \ldots, i_r \), \( d_{i_1,i_2,\ldots,i_r,k} \) is the subboundary of \( d_{i_1,i_2,\ldots,i_r} \), and

\[ d_{i_1,i_2,\ldots,i_r} \supset \bigcup_{k \neq i_1,i_2,\ldots,i_r} d_{i_1,i_2,\ldots,i_r,k} \]
When \( p = 3 \) as in Fig. A1, \( d_1 = CB, d_2 = AC, d_3 = AB \). The subboundary of \( d_1 \) is \( d_{12} = \{C\}, d_{13} = \{B\} \).

**Theorem A1.** If \( r_k(X) = p \), then model \( PR \) has a unique solution. Let \( b^* \) be the solution of equations

\[
\begin{align*}
\lambda' p b &= 1 \\
\lambda' p + X'Xb &= X'Y
\end{align*}
\]

If \( b^* \geq 0 \), then \( b^* \) is the solution of model \( PR \). If vector \( b^* \) has \( r \) negative components, that is, \( b^*_{i1} < 0, \ldots, b^*_{ir} < 0 \), then the solution is certainly on the boundary \( d_{i1} \cup d_{i2} \cup \cdots \cup d_{ir} \).

More detailed proof can be seen in Fang et al. (1982).

**References**


