# 22

# *Comparing Latent Class Factor Analysis with the Traditional Approach in Data Mining*

**Jay Magidson and Jeroen K. Vermunt**
*Statistical Innovations Inc., USA, and Tilburg University, The Netherlands*

**CONTENTS**



A major goal of data mining is to extract a small number of meaningful "factors" from a larger number of variables available on a database. While traditional factor analysis (FA) offers such a data reduction capability, it is severely limited in practice because it requires all variables to be continuous, and it uses the assumption of multivariate normality to justify a linear model. In this paper, we propose a general maximum likelihood alternative to FA that does not have the above limitations. It may be used to analyze combinations of dichotomous, nominal, ordinal, and count variables and uses appropriate distributions for each scale type. The approach utilizes a framework based on latent class (LC) modeling that hypothesizes categorical as opposed to continuous factors, each of which has a small number of discrete levels. One surprising result is that exploratory LC factor models are identified while traditional exploratory FA models are not identified without imposing a rotation.

## 22.1 Introduction

A major goal of data mining is to extract a relatively small number of meaningful "factors" from a larger number of variables available on a database. While traditional factor analysis (FA) offers such a data reduction capability, it is severely limited in practice for 4 reasons:

1. It requires all variables to be continuous.

2. It uses the assumption of multivariate normality to justify a linear model.

0-58488-344-8/04/$0.00+$1.50

3. It assumes that the underlying latent variables (factors) are measured on an interval or ratio scale.

4. Results are generally not unique – in order to interpret the solution users must select from among possible "rotations," each of which provides a somewhat different result.

Although justified only for continuous variables, FA is frequently used in practice with variables of other scale types including dichotomous, nominal, ordinal, and count variables. In such cases, the linearity assumption will generally be violated, as the true model will typically be nonlinear. In particular, when the observed variables are dichotomous, FA users have sometimes observed the occurrence of non-informative extraneous factors on which variables having a common skewness (i.e., similar marginal distributions) tend to have large factor loadings. It is possible that such factors serve as proxies for various nonlinearities in the model.

Even when the first 2 assumptions hold true, in the case that one or more *latent* variables is dichotomous, statistical inferences used in maximum likelihood FA are not valid as such tests assume that the factors are multivariate normal.

A promising alternative to FA, proposed by Magidson and Vermunt (2001) utilizes a framework based on latent class (LC) modeling. This latent class approach to factor analysis (LCFA) hypothesizes dichotomous or ordered categorical (ordinal) as opposed to continuous factors, and is especially suited for categorical variables. While this methodology resolves each of the 4 FA problems stated above, it has its own limitations. In particular, when used in the exploratory setting, the following limitations have been noted:

1. LCFA has primarily been applied in confirmatory applications involving a relatively small number of variables. Recent advances in computing power and the availability of new efficient algorithms suggest that LCFA may be applicable in larger exploratory settings, but this has not yet been tested.

2. The LCFA analogs to the "loadings" used in FA are given by log-linear parameters, which are not so easy to interpret.

In this paper, we use real data to compare LCFA with FA in situations where the assumptions from FA are violated. For simplicity, we have limited our current study to examples where the manifest variables are all dichotomous. To facilitate this comparison, we linearize the latent class model, transforming the log-linear effects to linearized parameters comparable to traditional loadings used in FA. Two data sets are used for this comparison. The first utilizes data analyzed previously by LCFA (Magidson and Vermunt, 2003) which yields results that are clearly nonlinear. The second involves 19 dichotomous indicators from the Myers-Briggs Type Indicator, designed to measure 2 latent dimensions of personality, hypothesized by Karl Jung to be dichotomous.

The LCFA model is described in section 22.2 along with a brief history of LC models. Section 22.3 presents results from 2 data examples, where the Latent GOLD

computer program was used to estimate the LC models. The results are summarized in section 22.4.

## 22.2 The Basic LC Factor Model

The LC Factor model was originally proposed for use with nominal manifest and dichotomous latent variables in various confirmatory applications (Goodman, 1974). This model was extended for use with ordinal *latent* variables and for *manifest* variables of differing scale types – dichotomous, ordinal, continuous, and count variables – by Vermunt and Magidson (2000). A *basic* LC factor model consisting of $K$ mutually independent dichotomous factors was proposed by Magidson and Vermunt (2001) for general exploratory applications. For the expository purposes of this paper, we limit to applications of this exploratory model. For other applications involving more complex LCFA models see Magidson and Vermunt (2003).

Let $\theta_k$ denote a value of one of the $K$ dichotomous latent variables. Without loss of generality we assume that $\theta_k$ can take on two values, 0 and 1. Let $y_j$ denote a value on one of the $J$ observed variables. The most common parameterization of the basic LC factor model is in terms of unconditional and conditional probabilities. For example, for 4 nominal variables, a basic 2-factor LC model can be expressed in terms of the joint probability $P(\theta_1,\theta_2,y_1,y_2,y_3,y_4)$:

$$\begin{aligned} P(\theta_1,\theta_2,y_1,y_2,y_3,y_4) = {} & P(\theta_1,\theta_2)P(y_1|\theta_1,\theta_2)P(y_2|\theta_1,\theta_2) \\ & P(y_3|\theta_1,\theta_2)P(y_4|\theta_1,\theta_2), \end{aligned} \tag{22.1}$$

where the conditional probability parameters are restricted by logit models. More precisely, the conditional probability for manifest variable $j$ is assumed to be equal to

$$P(y_j|\theta_1,\theta_2) = \frac{\exp(\beta_{j0y_j} + \beta_{j1y_j}\theta_1 + \beta_{j2y_j}\theta_2)}{\sum_{y_j}\exp(\beta_{j0y_j} + \beta_{j1y_j}\theta_1 + \beta_{j2y_j}\theta_2)}. \tag{22.2}$$

For the basic factor LC model, the latent variables are assumed to be independent of one another. Thus, we have the following additional constraint:

$$P(\theta_1,\theta_2) = P(\theta_1)P(\theta_2). \tag{22.3}$$

The constraints of the type in equation (22.2) restrict the conditional response probabilities in a manner similar to traditional FA by excluding the higher-order interaction terms involving $\theta_1$ and $\theta_2$. The $\beta$ parameters can be viewed as category-specific "loadings" on the factor concerned, expressed as log-linear parameters. Note that one identifying constraint has to be imposed on each set of $\beta$ parameters.

If variable $j$ were instead ordinal or dichotomous, equation (22.2) becomes

$$P(y_j|\theta_1,\theta_2) = \frac{\exp(\beta_{j0y_j} + \beta_{j1}y_j\theta_1 + \beta_{j2}y_j\theta_2)}{\sum_{y_j}\exp(\beta_{j0y_j} + \beta_{j1}y_j\theta_1 + \beta_{j2}y_j\theta_2)}. \tag{22.4}$$

in which case a single loading for variable $j$ on each of the 2 factors is given by $\beta_{j1}$ and $\beta_{j2}$, respectively.

More generally, let $\theta$ denote a vector of $K$ latent variables and $\mathbf{y}$ a vector of $J$ observed variables. Then the model becomes:

$$f(\theta,\mathbf{y}) = f(\theta)f(\mathbf{y}|\theta) = f(\theta)\prod_{j=1}^{J} f(y_j|\theta) \tag{22.5}$$

where $f(\theta,\mathbf{y})$ denotes the joint probability density of the latent and manifest variables, $f(\theta)$ the unconditional latent probabilities, and $f(y_j|\theta)$ the conditional density for variable $j$ given a person's latent scores. The primary model assumption in equation (22.5) is that the $J$ observed variables are independent of each other given the latent variables. That is, as in traditional FA, the latent variables explain all of the associations among the observed variables.

The conditional means of each manifest variable are restricted by regression type constraints; that is, by a regression model from the generalized linear modelingfamily. The following distributions and transformations are used:

| Scale type | Distribution $f(y_j\|\theta)$ | Transformation $g(.)$ |
|---|---|---|
| dichotomous | binomial | logit |
| nominal | multinomial | logit |
| ordinal | multinomial | restricted logit |
| count | Poisson | log |
| continuous | normal | identity |

In equations (22.2) and (22.4), we gave the form of the regression models for dichotomous, nominal, and ordinal variables.

Parameters can be estimated by maximum likelihood using EM or Newton-Raphson algorithms, or combinations of the two. Maximum likelihood estimation of the LCFA model is implemented in the Latent GOLD program (Vermunt and Magidson, 2000).

## 22.3 Examples

In this section we compare results obtained from the latent class factor model with the traditional linear factor model.

### 22.3.1 Rater Agreement

For our first example we factor analyze ratings made by 7 pathologists, each of whom classified 118 slides as to the presence or absence of carcinoma in the uterine cervix (Landis and Koch, 1977). Agresti (2002), using traditional LC models to analyze these data, found that a 2-class solution does not provide an adequate fit to these data. Using the LCFA framework, Magidson and Vermunt (2003) confirmed that a

| | Factor $\theta_1$ = 1 (True -) | | Factor $\theta_2$ = 0 (True +) | |
|---|---|---|---|---|
| | Factor $\theta_1$ | | Factor $\theta_2$ | |
| | 1 | 0 | 1 | 0 |
| Rater | 0.35 | 0.18 | 0.31 | 0.16 |
| F | | | | |
| - | 1.00 | 0.99 | **0.80** | 0.11 |
| + | 0.00 | 0.01 | 0.20 | 0.89 |
| D | | | | |
| - | 1.00 | 0.98 | **0.61** | 0.11 |
| + | 0.00 | 0.02 | 0.39 | 0.89 |
| C | | | | |
| - | 1.00 | 1.00 | 0.22 | 0.14 |
| + | 0.00 | 0.00 | 0.78 | 0.86 |
| A | | | | |
| - | 0.94 | 0.59 | 0.01 | 0.00 |
| + | 0.06 | **0.41** | 0.99 | 1.00 |
| G | | | | |
| - | 0.99 | 0.46 | 0.01 | 0.00 |
| + | 0.01 | **0.54** | 0.99 | 1.00 |
| E | | | | |
| - | 0.94 | 0.28 | 0.03 | 0.00 |
| + | 0.06 | **0.72** | 0.97 | 1.00 |
| B | | | | |
| - | 0.87 | 0.01 | 0.03 | 0.00 |
| + | 0.13 | **0.99** | 0.97 | 1.00 |

**TABLE 22.1**
Estimates of the unconditional and conditional probabilities obtained from the 2-factor LC Model.

single dichotomous factor (equivalent to a 2-class LC model) did not fit the data but that a basic 2-factor LCFA model provides a good fit.

Table 22.1 presents the results of the 2-factor model in terms of the conditional probabilities. These results suggest that factor 1 distinguishes between slides that are "true positive" or "true negative" for cancer. Factor 2 is a nuisance factor, which suggests that some pathologists bias their ratings in the direction of a "false +" error while others exhibit a bias towards "false -" error. Overall, these results demonstrate the richness of the LCFA model to extract meaningful information from these data. Valuable information includes an indication of which slides are positive for carcinoma, as well as estimates of "false +" and "false -" error for each rater.

The left-most columns of Table 22.2 list the estimates of the log-linear parameters for these data. Although the probability estimates in Table 22.1 are derived from

| | Log-linear | | Communalities based on | | Linearized model | | |
|---|---|---|---|---|---|---|---|
| Rater | $\theta_1$ | $\theta_2$ | Linear terms only | | Total | $\theta_1$ | $\theta_2$ |
| F | 7.2 | 3.4 | 0.45 | 0.60 | 0.53 | 0.38 | 0.40 |
| D | 6.0 | 2.6 | 0.47 | 0.54 | 0.62 | 0.26 | 0.26 |
| C | 7.2 | 0.5 | 0.68 | 0.68 | 0.82 | 0.04 | 0.04 |
| A | 7.7 | 2.4 | 0.72 | 0.75 | 0.82 | 0.18 | -0.16 |
| G | 10.1 | 5.2 | 0.76 | 0.82 | 0.82 | 0.27 | -0.25 |
| E | 6.4 | 3.8 | 0.65 | 0.75 | 0.72 | 0.35 | -0.31 |
| B | 5.3 | 6.3 | 0.59 | 0.76 | 0.60 | 0.47 | -0.42 |

**TABLE 22.2**
Log-linear and Linearized Parameter Estimates for the 2-factor LC Model.

these quantities (recall equation 22.2), the log-linear estimates are not as easy to interpret as the probabilities.*

Traditional factor analysis fails to capture the differential biases among the raters. Using the traditional rule of choosing the number of factors to be equal to the number of eigenvalues greater than 1 yields only a single factor. (The largest eigenvalue is 4.57, followed by 0.89 for the second largest.) For purposes of comparison with the LCFA solution, we fit a 2-factor model using maximum likelihood for estimation.

Table 22.3 shows the results obtained from both varimax and quartimax rotations. The substantial differences between these loadings is not a reliable method for extracting meaningful information from these data.

The right-most columns of Table 22.2 present results from a linearization of the LCFA model using the following equation to obtain "linearized loadings" for each variable $j$:

$$E(y_j|\theta_1,\theta_2) = \rho_{j0} + \rho_{j1}\theta_1 + \rho_{j2}\theta_2 + \rho_{j12}\theta_1\theta_2. \tag{22.6}$$

These 22.3 loadings have clear meanings in terms of the magnitude of validity and bias for each rater. They have been used to sort the raters according to the magnitude and direction of bias. The log-linear loadings do not provide such clear information.

The loading on $\theta_1$ corresponds to a measure of validity of the ratings. Raters C, A, and G who have the highest loadings on the first linearized factor show the highest level of agreement among all raters (Magidson and Vermunt, 2003). The loading on $\theta_2$ relates to the magnitude of bias and the loading on $\theta_1\theta_2$ indicates the direction of the bias. For example, from Table 22.1 we saw that raters F and B show the most

*For example, the log-linear effect of A on $\theta_2$, a measure of the validity of the ratings of pathologist A, is a single quantity, exp(7.74)=2,298. This means that among those slides at level 1 of $\theta_2$, the odds of rater A classifying a "true +" slide as "+" is 2,298 times as high as classifying a "true -" slide as "+". Similarly, among those slides at level 0 of $\theta_2$, this expected odds ratio is also 2,298. The linear measure of effect is easier to interpret, but is not the same for both types of slides. For slides at level 1 of $\theta_2$, the probability of classifying a "true +" slide as "+" is .94 higher (.99-.06=.93), while for slides at level 0 of $\theta_2$, it is .59 higher (1.00 - .41=.59), a markedly different quantity.

| Rater | Comm-unalities | Varimax Rotation Factor $\theta_1$ | Varimax Rotation Factor $\theta_2$ | Quartimax Rotation Factor $\theta_1$ | Quartimax Rotation Factor $\theta_2$ |
|---|---|---|---|---|---|
| F | 0.49 | 0.23 | 0.66 | 0.55 | 0.43 |
| D | 0.60 | 0.29 | 0.72 | 0.63 | 0.45 |
| C | 0.62 | 0.55 | 0.56 | 0.77 | 0.18 |
| A | 0.73 | 0.71 | 0.48 | 0.85 | 0.03 |
| G | 0.86 | 0.83 | 0.42 | 0.92 | -0.09 |
| E | 0.78 | 0.82 | 0.31 | 0.86 | -0.18 |
| B | 0.69 | 0.80 | 0.24 | 0.80 | -0.22 |

**TABLE 22.3**
Results Obtained from a 2-factor Solution from Traditional Factor Analysis.

bias, F in the direction of "false -" ratings and B in the direction of "false +". The magnitude of the loadings on the nonlinear term is highest for these 2 raters, one occurring as "+," the other as "-."

Table 22.2 also lists the communalities for each rater, and decomposes these into linear and nonlinear portions (the "total" column includes the sum of the linear and nonlinear portions). The linear portion is the part accounted for by $\rho_{j1}\theta_1 + \rho_{j2}\theta_2$, and the nonlinear part concerns the factor interaction $\rho_{j12}\theta_1\theta_2$. Note the substantial amount of nonlinear variation that is picked up by the LCFA model. For comparison, the right-most column of Table 22.4 provides the communalities obtained from the FA model.

### 22.3.2 MBTI Personality Items

Our second example consists of 19 dichotomous items from the Myers-Briggs Type Indicator (MBTI) – 7 indicators of the Sensing-iNtuition dimension, and 12 indicators of the Thinking-Feeling personality dimension. These items are designed to measure 2 hypothetical personality dimensions, which were posited by Carl Jung to be latent dichotomies.

The log-likelihood values obtained from fitting 0, 1, 2, 3 LC factor models are summarized in Table 22.4. Strict adherence to the BIC, AIC, CAIC or similar criterion suggest that more than 2 latent factors are required to fit these data due to violations of the local independence assumption. This is due to similar wording[†] used in several of the S-N items and similar wording used in some of the T-F items.

[†]For example, in a 3-factor solution, all loadings on the third factor are small except those for S-N items S09 and S73. Both of these items ask the respondent to express a preference between "practical" and a second alternative (for item S09, "ingenious"; for item S73, "innovative").

| Model | Log-likelihood (LL) | % of LL Explained | Time in seconds ‡ |
|---|---|---|---|
| 0-Factor | 72804 | 0.00 | 1 |
| 1-Factor | 55472 | 0.24 | 5 |
| 2-Factor | 46498 | 0.36 | 11 |
| 3-Factor | 43328 | 0.40 | 27 |

**TABLE 22.4**
Results of Estimating LC Factor Models.

In such cases, additional association between these items exists which is not explainable by the general S-N (T-F) factor. For our current purpose, we ignore these local dependencies and present results of the 2-factor model.

The right-most column of Table 22.2 shows that estimation time is not a problem. Estimation of the 3-factor model using the Latent GOLD computer program took only 27 seconds on a 2000 Megahertz computer.

In contrast to our first example, the decomposition of communalities in the right-most columns of Table 22.5 shows that a linear model can approximate the LCFA model here quite well. Only for a couple of items is the total communality not explained to 2 decimal places by the linear terms only. The left-most columns of Table 22.5 compares the log-linear and linearized "loadings" for each variable. The fact that the latter numbers are bounded between -1 and +1 offers easier interpretation.

The traditional FA model also does better here than the first example. The first four eigenvalues turn out to be 4.4, 2.8, 1.1 and 0.9. For comparability to the LC solution, Table 22.6 presents the loadings for the 2-factor solution under Varimax and Quartimax rotations. Unlike the first example where the corresponding loadings showed considerable differences, these two sets of loadings are quite similar. The results are also similar to the linearized loadings obtained from the LCFA solution.

The right-most column of Table 22.6 shows that the communalities obtained from FA are quite similar to those obtained from LCFA. Generally speaking, these communalities are somewhat higher than those for LCFA, especially for items S27, S44, and S67 (highlighted in bold).

Figure 22.1 displays the 2-factor bi-plot for these data (see Magidson and Vermunt, 2001). The plot shows how clearly differentiated the S-N items are from the T-F items on the 2-factors. The 7 S-N items are displayed along the vertical dimension of the plot, which is associated with factor 2, while the T-F items are displayed along the horizontal dimension, which is associated with factor 1. This display turns out to be very similar to the traditional FA loadings plot for these data. The advantage of this type of display becomes especially evident when nominal variables are included among the items.

| Item | Log-linear | | Linearized | | Communalities based on | |
|---|---|---|---|---|---|---|
| | $\theta_1$ | $\theta_2$ | $\theta_1$ | $\theta_2$ | Linear terms only | Total |
| S02 | 0.03 | **-1.51** | -0.01 | **-0.61** | 0.37 | 0.37 |
| S09 | 0.01 | **-1.16** | 0.00 | **-0.50** | 0.25 | 0.25 |
| S27 | -0.03 | **1.46** | 0.01 | **0.55** | 0.30 | 0.30 |
| S34 | 0.07 | **-1.08** | -0.03 | **-0.45** | 0.21 | 0.21 |
| S44 | 0.11 | **1.13** | -0.04 | **0.47** | 0.22 | 0.22 |
| S67 | 0.06 | **1.54** | -0.02 | **0.53** | 0.28 | 0.28 |
| S73 | 0.01 | **-1.05** | 0.00 | **-0.46** | 0.21 | 0.21 |
| T06 | **-1.01** | 0.53 | **0.43** | 0.19 | 0.22 | 0.22 |
| T29 | **-1.03** | 0.59 | **0.44** | 0.20 | 0.23 | 0.23 |
| T31 | **1.23** | -0.47 | **-0.52** | -0.15 | 0.29 | 0.29 |
| T35 | **1.42** | -0.29 | **-0.55** | -0.09 | 0.31 | 0.32 |
| T49 | **-1.05** | 0.65 | **0.44** | 0.22 | 0.24 | 0.25 |
| T51 | **-1.32** | 0.30 | **0.53** | 0.09 | 0.29 | 0.29 |
| T53 | **-1.40** | 0.77 | **0.56** | 0.22 | 0.36 | 0.36 |
| T58 | **1.46** | -0.12 | **-0.62** | -0.03 | 0.38 | 0.38 |
| T66 | **1.23** | -0.27 | **-0.54** | -0.09 | 0.30 | 0.30 |
| T70 | **-1.07** | 0.61 | **0.43** | 0.19 | 0.22 | 0.23 |
| T75 | **1.01** | -0.39 | **-0.45** | -0.14 | 0.22 | 0.22 |
| T87 | **1.17** | -0.45 | **-0.50** | -0.15 | 0.28 | 0.28 |

**TABLE 22.5**
Log-linear and Linearized Parameter Estimates and Communalities for the 2-Factor LC Model as Applied to 19 MBTI items.

## 22.4 Conclusion

In this study, we compared LCFA with FA in 2 cases where the assumptions from FA were violated. In one case, the resulting linear factor model obtained from FA provided results that were quite similar to those obtained from LCFA even though the factors were taken to be dichotomous in the LCFA model. In this case, decomposition of the LCFA solution into linear and nonlinear portions suggested that the systematic portion of the results was primarily linear, and the linearized LCFA solution was quite similar to the FA solution. However, the LCFA model was able to identify pairs and small groups of items that have similar wording because of some violations of the assumption of local independence.

In the second case, LCFA results suggested that the model contained a sizeable nonlinear component, and in this case the FA result was unable to capture differential

| Item | Quartimax Rotated Factor Matrix | | Varimax Rotated Factor Matrix | | Comm- |
|---|---|---|---|---|---|
| | Factor | | Factor | | |
| | 1 | 2 | 1 | 2 | unalities |
| S02 | 0.08 | **-0.63** | 0.06 | **-0.63** | 0.40 |
| S09 | 0.07 | **-0.50** | 0.06 | **-0.50** | 0.26 |
| S27 | -0.06 | **0.62** | -0.05 | **0.62** | **0.38** |
| S34 | 0.07 | **-0.46** | 0.06 | **-0.46** | 0.22 |
| S44 | -0.02 | **0.55** | 0.00 | **0.55** | **0.30** |
| S67 | -0.02 | **0.64** | -0.01 | **0.64** | **0.41** |
| S73 | 0.06 | **-0.46** | 0.05 | **-0.46** | 0.21 |
| T06 | **-0.49** | 0.09 | **-0.49** | 0.10 | 0.25 |
| T29 | **-0.49** | 0.10 | **-0.49** | 0.11 | 0.25 |
| T31 | **0.56** | -0.04 | **0.56** | -0.05 | 0.32 |
| T35 | **0.58** | 0.05 | **0.58** | 0.04 | 0.34 |
| T49 | **-0.50** | 0.13 | **-0.50** | 0.15 | 0.27 |
| T51 | **-0.57** | -0.03 | **-0.57** | -0.02 | 0.33 |
| T53 | **-0.61** | 0.09 | **-0.61** | 0.10 | 0.38 |
| T58 | **0.64** | 0.11 | **0.64** | 0.10 | 0.42 |
| T66 | **0.58** | 0.05 | **0.58** | 0.03 | 0.33 |
| T70 | **-0.49** | 0.10 | **-0.49** | 0.11 | 0.25 |
| T75 | **0.50** | -0.03 | **0.50** | -0.04 | 0.25 |
| T87 | **0.55** | -0.04 | **0.55** | -0.05 | 0.30 |

**TABLE 22.6**
Results from Traditional Factor Analysis of the 19 MBTI items.

biases between the raters. Even when a second factor was included in the model, no meaningful interpretation of this second factor was possible, and the loadings from 2 different rotations yielded very different solutions.

Overall, the results suggest improved interpretations from the LCFA approach, especially in cases where the nonlinear terms represent a significant source of variation. This is due to the increased sensitivity of the LCFA approach to all kinds of associations among the variables, not being limited as the FA model to the explanation of simple correlations.

The linearized LCFA parameters produced improved interpretation, but in the nonlinear example, a third (nonlinear) component model was needed in order to extract all of the meaning from the results. This current investigation was limited to 2 dichotomous factors. With 3 or more dichotomous factors, in addition to each 2-way interaction, additional loadings associated with components for each higher-order interaction would also be necessary. Moreover, for factors containing 3 or more levels, additional terms are required. Further research is needed to explore these issues in practice.

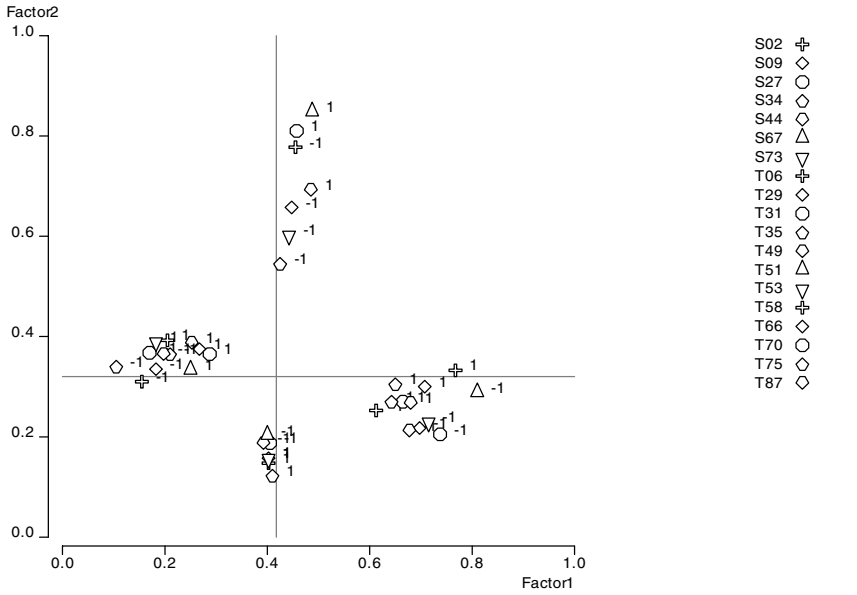


**FIGURE 22.1**
**2-factor bi-plot.**

## References

Agresti, A. (2002). *Categorical Data Analysis*. Second Edition. New York: Wiley.

Goodman, L.A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models, *Biometrika*, 61, 215-231.

Landis, J.R. and Koch, G.G. (1977). The measurement of observer agreement for categorical data, *Biometrics*, 33, 159-174.

Magidson, J. and Vermunt, J.K. (2001). Latent class factor and cluster models, bi-plots and related graphical displays, *Sociological Methodology*, 31, 223-264.

Magidson, J. and Vermunt, J.K. (2003). Latent class models, chapter in D. Kaplan (editor), *Handbook of Quantitative Methods in Social Science Research*, Sage Publications, Newbury Park, CA.

Vermunt, J.K. and Magidson, J. (2000). *Latent GOLD 2.0 User's Guide*. Belmont, MA: Statistical Innovations Inc.