A Bootstrap for Latent Class Models to Determine the Appropriate Number of Latent Classes

Suzanne Winsberg
IRCAM
1 Place Igor Stravinsky
Paris, 75004, France
winsberg@ircam.fr

Geert de Soete
University of Ghent, Academic Computing Center
Krijgslaan 281, Building S9
B-9000 Ghent, Belgium
desoete@rug.ac.be

1. Introduction

Latent class formulations, or more general mixture distribution approaches, have been explored in many contexts. In all of these applications, latent class modeling has been shown to be a useful technique for capturing systematic group differences in a parsimonious way. A problem encountered in the latent class approach is to determine the appropriate number of latent classes. A Monte Carlo procedure suggested by Hope (1968) has been used to address this problem. We propose another approach using the bootstrap. The bootstrap, introduced by Efron (1979), is a resampling technique in which B random samples are drawn with replacement from the observed values. Our technique can be used in lieu of or in addition to the Hope test. This bootstrap approach may be used in any latent class context; however, we illustrate the approach in the multidimensional scaling context using CLASCAL, a latent class MDS model, (Winsberg and De Soete, 1993). Multidimensional scaling, MDS, may be defined broadly as a set of multivariate models and methods for representing objects as points in a multidimensional space of low dimension, given pairwise dissimilarity measures for the objects. In addition to the bootstrap procedure for fixing the number of latent classes, we propose two bootstrap methods to assess the standard errors of the MDS model parameters, namely the object coordinates in the multidimensional space.

2. Bootstrapping the Appropriate Number of Classes

Consider a data set with pairwise dissimilarity judgements for n objects, from N subjects, or sources. To test the postulate that T classes are sufficient to account for group differences in the data, we propose to conduct a bootstrap analysis for T classes, for (T+1) classes and for (T+1) classes. For each analysis draw B bootstrap samples of size N; that is, draw B samples with replacement of size N. Consider the analysis for T classes. Construct the N X N matrix in which the ikth element is 1 - (the number of times the ith subject is a member of the same class as the kth subject divided by the number of times both the ith and kth subject appear in any of the bootstrap samples). This matrix, containing a measure of dissimilarity for each pair of subjects, should then be analyzed with an average link hierarchical clustering algorithm. If indeed there are T classes, the clustering analysis will yield a tree that shows T distinct groups.
3. An MDS Application: Bootstrapping Standard Errors for the Model Parameters

Once the appropriate number of latent classes has been determined, and the MDS model that best fits the data has been selected, standard errors for the object coordinates may be determined using one of the two following bootstrap techniques: the first, denoted the naive bootstrap, consists of drawing with replacement B bootstrap samples, of size N; the second is conditional on $\mathbf{\Lambda}$, the vector consisting of all unconditional class membership probabilities. Two types of bootstrap are proposed, as in some cases one may require that the results of the bootstrap be conditional on class size, while in other cases that may not be the case. To determine standard errors one must combine the results obtained for each bootstrap sample. Attention must be paid to constraints imposed on the analyses to identify the solution, when combining the results.

4. An Artificial and a Real Data Example

We have analyzed artificial data, comprising pairwise dissimilarity judgements from 16 subjects on 20 rectangles, differing as to shape and area. These artificial data are based on real data we have previously analyzed, (Winsberg and DeSoete, 1996). Using our bootstrap approach we were correctly able to determine the number of latent classes. For this example Hope’s technique failed to do so. Moreover, we were able to replicate standard errors found from a Monte Carlo analysis of 40 samples generated from the model for the artificial data using normal error. We have also analyzed real data comprising pairwise dissimilarity data from 15 subjects judging 9 simulated concert hall configurations varying in reverberation time and clarity. The standard errors for the object coordinates revealed that certain concert hall configurations were located with much greater precision than others.

In conclusion these bootstrap techniques are useful both for deciding how many latent classes should be retained, and for obtaining standard errors for the parameters.

REFERENCES


RÉSUMÉ

Nous proposons un bootstrap adapté aux modèles de classes latentes qui permet de déterminer le nombre de classes suffisant pour expliquer les différences systématiques entre les groupes de sujets échantillonnés.