On likelihood ratio tests for dimensionality selection

Yoshio Takane
McGill University

1. Let me start by saying that this is primarily an expository talk; there is nothing radically new, but it will very informative, particularly because so many people (including myself) have made essentially the same mistake repeatedly in the past.
2. There are a number of models in psychometrics having dimensional structure. Here I have several examples of such models; MDS, FA/PCA, … and so on. These models require judicious choice of dimensionality.
3. When we have two nested models (one is a special case of the other), the LR statistic between the two models "usually" follows an asymptotic chi-square distribution (with df's equal to the difference in the effective number of parameters in the two models) under the hypothesis that the more restricted model is correct. Here Model B is nested within A, which in turn is nested within the saturated model. The LR statistic between any of these two models is asymptotically chi-square ((1) versus (2) with *a* df, (2) vs (3) with *b* df, and then (1) vs (3) χ^2 with $a+b$ df (by the reproducibility of the χ^2 distribution).
4. So we are tempted to use the same scheme to test the dimensionality. (The r-dimensional model is a special case of the $(r+1)$-dimensional model.) Unfortunately, this LR statistic is not likely to follow an asymptotic chi-square distribution. Here, (1) vs (3) is asymptotic χ^2, provided that (3) is correct but (4) is not. However, neither (1) vs (2) (unless (3) is incorrect) nor (2) vs (3) is likely to follow asymptotic χ^2.
5. So what's so special in this situation? To explain we introduce the notion of local regularity due to Shapiro (1986). "A point θ_0 in the parameter space Ω is said to be locally regular if the rank of the Jacobian matrix stays the same for all θ in a neighborhood θ_0, where the Jacobian matrix is defined as the matrix of the derivatives of model predictions wrt model parameters. This definition also implies that θ_0 lies in the interior (not on the boundary) of Ω, so that there is a neighborhood of θ_0 in Ω. When the local regularity is violated the LR statistic is not guaranteed to follow an asymptotic chi-square.
6. Let us illustrate. Here we have 4 stimuli represented in a 2-dimensional Euclidean space. This is the matrix of stimulus coordinates, where we fixed these elements to zero to eliminate the rotational and translational indeterminacies in the Euclidean distance model. So the parameter vector has 5 elements, and the model vector has 6 elements.
7. The Jacobian matrix looks like this. (It is a 5 by 6 matrix obtained by differentiating each element of the model vector wrt each element of the parameter vector.) δ_{ija} is as defined here, where we assumed that no d_{ij}'s are zero.
8. Let's look at the following two cases. In the first case we are testing if this particular element (x_{32}) is 0 or not. The rank(J) is 5 x_{32} = 0, and that stays the same for all values of x_{32} near 0, so that the local regularity condition holds, and the LR statistic (λ_c) between H_0 and H_1 has an asymptotic chi-square distribution with 1

df. Now consider the second case. We are testing if the model is 1-dimensional or 2-dimensional. The null hypo. (H_0) states that $x_{32} = x_{42} = 0$. The rank(J) is 3 under H_0, but it suddenly becomes 5 as soon as it departs from this point (that is, under H_1). The local regularity is violated, and the LR statistic (λ_c) is not guaranteed to follow an asymptotic chi-square dist. (Note, however, the LR statistic (λ_s) defined between the saturated model and the 1-dimensional model follows an asymptotic chi-square, unless the 0-dimensional model is also true.)

9. I am putting the wording carefully. A violation of local regularity does not imply that λ_c is never asymptotically chi-square. It is just that it is not guaranteed to follow an asymptotic chi-square, because the local regularity is a sufficient, but not a necessary, condition for asymptotic chi-square. This means that when this condition is violated, the asymptotic distribution of λ_c has to be examined in each specific situation. It is also the case that the theory presented does not tell us how much the distribution of λ_c deviates from a chi-square distribution.
10. So a Monte Carlo study is in order. We again consider an MDS situation where we have 10 stimuli represented in a two-dimensional Euclidean space. (This is the true model.) The data were generated in tetradic form according to a prescribed conf., and MAXSCAL-4, a maximum likelihood MDS program, was used for parameter estimation. Both λ_c (the comparison between the 2- and 3-dimensional solutions) and λ_s (the comparison between the 2-dimensional and the saturated models) were calculated for each of the 100 data sets.
11. Here are q-q plots of the LR statistics against the theoretical distributions. Left panels are for λ_c and the right λ_s for 4 replicated observations (top) and 20 replicated observations (bottom). While λ_s converges nicely to a theoretical distribution rather quickly, that is not the case for λ_c.
12. This persists when the replication size was increased to 500 and 2000, although the departure from the chi-square dist. is not that big in this particular case.
13. Conclusion: 1). Use λ_s as much as possible. (You first compare the 0- or 1-dimensional model against the saturated. If it is significant, increase the dimension by one, and compare that against the saturated, and so on, until no significant difference is found.) 2). However, there are models for which no saturated model exists (e.g., RA, parametric mixture models, NN models, etc.) 3). May have to resort to some numerical methods such as permutation tests, parametric Bootstrap, etc. 4). There are attempts to find the true dist of λ_c. 5). There are also cases the true dist. of λ_c is known.