The Impact of Model Misspecification in Clustered and Continuous Growth Modeling

Daniel J. Bauer
Odum Institute for Research in Social Science
The University of North Carolina at Chapel Hill

Patrick J. Curran
L.L. Thurstone Psychometric Laboratory
Department of Psychology
The University of North Carolina at Chapel Hill

A key focus of developmental science is the study of individual differences in developmental trajectories. Two types of models are often used for this purpose:

- Latent curve models (LCM) – Estimate quantitative variations in continuously distributed trajectory parameters.
- Growth mixture models (GMM) – Estimate latent classes of individuals whose trajectories differ in qualitatively important ways.

The relationship between these models can be clarified by writing the density function for the GMM as

$$g(y \mid \pi, \theta) = \sum_{k=1}^{K} \pi_k f_k(y \mid \theta_k)$$

where π_k represent the proportion of cases from class k, and θ_k are the growth parameters that define the trajectory f_k for each class k. This model reduces to the LCM when $K = 1$.

This poster investigates several critical issues raised by these alternative models:

- Can data screening guide the selection of an LCM or GMM model?
- Can model misspecification be detected using traditional fit statistics?
- How will model misspecification effect the estimation and interpretation of the model parameters?

Method

<u>Design Factors</u>

- The true number of growth functions in the population (K = 1 or 3)
- The model estimated for the data (LCM or GMM)
- The type of growth model fit to the data (Unconditional or Conditional)

<u>Data Generation</u>

- Multiclass (K = 3) and Single Class (K = 1) data were generated with Mplus.
    - The Multiclass data sets both combined an equal number of cases from each of the three trajectories displayed in Figure 1.
    - The Single Class data was generated only from the trajectory of Group 2.
- Each set included 6000 cases so the results would reflect asymptotic behavior.
- Two time-invariant predictors (X1-X2) were generated so that their relations to the trajectory parameters would vary across classes.

Results

<u>Data Screening</u>

- Histograms of Y1-Y4 are presented in Figure 2 for the Multiclass data.
    - The distributions appear normal at all time points except Y4.
- Histograms of individual parameter estimates (generated by case-wise OLS) for the Multiclass data are presented in Figure 3.

    - The distributions appear normal.
- Scatterplots of Y1-Y4 and the OLS parameters (not shown) also appeared bivariate normal.

Model Fit Diagnostics

- Good fit was obtained for both correct and misspecified LCM models.
    - LCM of Single Class data ($K = 1$)
        - Unconditional Model: $\chi^2(1) = .23$, p = .63; RMSEA = .00
        - Conditional Model: $\chi^2(3) = 1.42$, p = .70; RMSEA = .00
    - LCM of Multiclass data ($K = 3$)
        - Unconditional Model: $\chi^2(1) = .13$, p = .72; RMSEA = .00
        - Conditional Model: $\chi^2(3) = 1.68$, p = .64; RMSEA = .00
- GMM models ($K = 2$ & 3) fit to the single class data consistently iterated to a solution in which the "extra" groups had estimated sample sizes of 0.
- Fit statistics (BIC and AIC) for the GMM models fit to the Multiclass data are presented in Figure 4.
    - BIC & AIC suggested 2 rather than 3 groups for the unconditional models.
    - Discrimination of the 3 groups is improved with the inclusion of predictors.

Impacts of Misspecification

- Incorrectly specifying too many groups in a GMM analysis had no adverse effects on parameter estimates, as the "extra" groups had zero estimated members.
- Figure 5 presents the mean trajectories estimated for the Multiclass data when too few groups were specified.
    - The LCM ($K = 1$) model essentially averaged the three trajectory classes
    - The GMM ($K = 2$) model essentially merged groups 1 and 2
- In the conditional models, the effects of the predictors on the growth factors were similarly averaged as the groups were combined.

Conclusions

- Can data screening guide the selection of an LCM or GMM model?
    - Neither the distributions of the observed variables nor those of the growth parameters conveyed the presence of multiple trajectory classes.
    - Data screening may be more informative as the trajectory classes become more distinctive and the variance within classes decreases.
- Can model misspecification be detected using traditional fit statistics?
    - Misspecifying an LCM model for Multiclass data was not detectable using traditional model fit statistics.
    - BIC & AIC did not always lead to selection of the correct number of groups.
        - Including predictors in model helped to discriminate the groups.
- How will model misspecification effect the estimation and interpretation of the model parameters?
    - Estimating a GMM with more classes than necessary had no harmful effects, and pointed toward simplifying the model to the correct form.
        - The effect of this type of misspecification would probably be greater under less ideal conditions (e.g, when single class data are skewed).
    - When too few groups were specified, the most similar groups collapsed together, obscuring their distinctive trajectories and relations to predictors.

Figure 1. Trajectories used for Data Generation

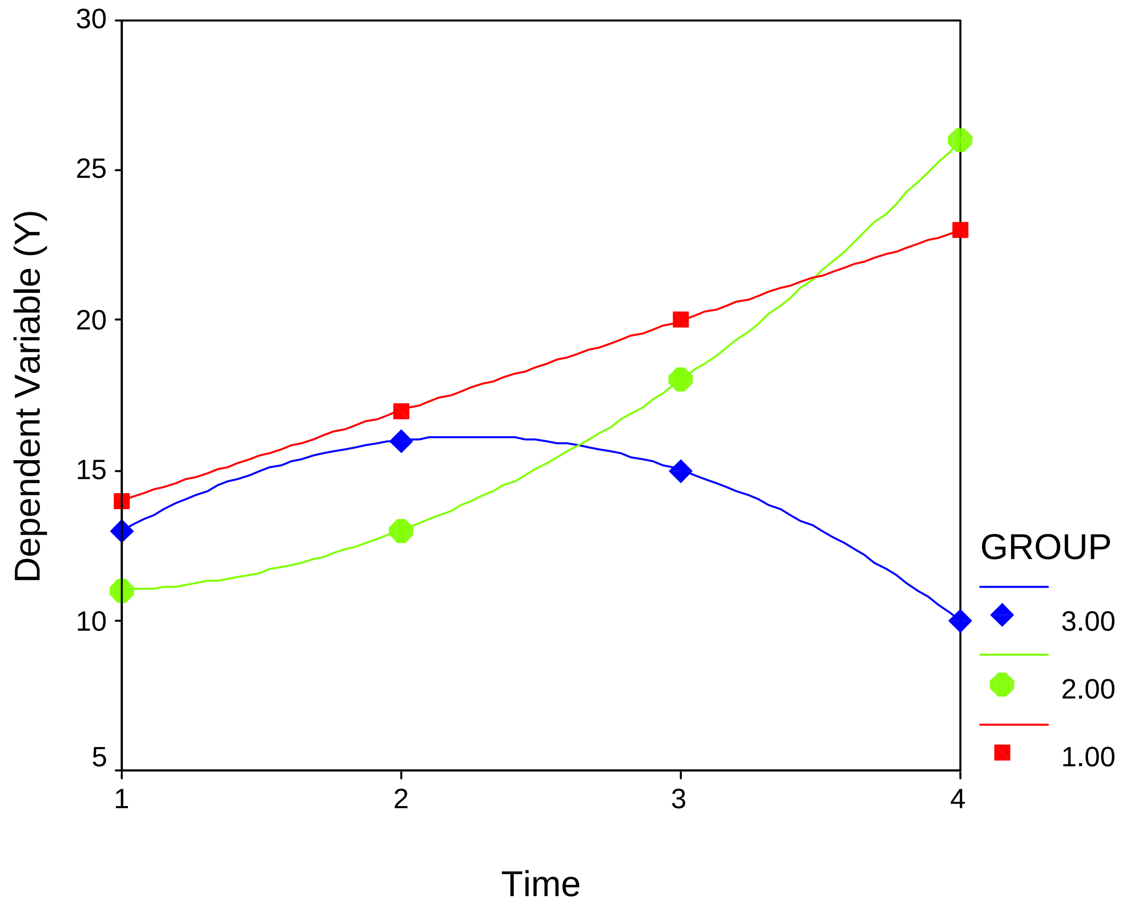

Figure 2. Univariate Distributions of Y1–Y4 For Multiclass Data

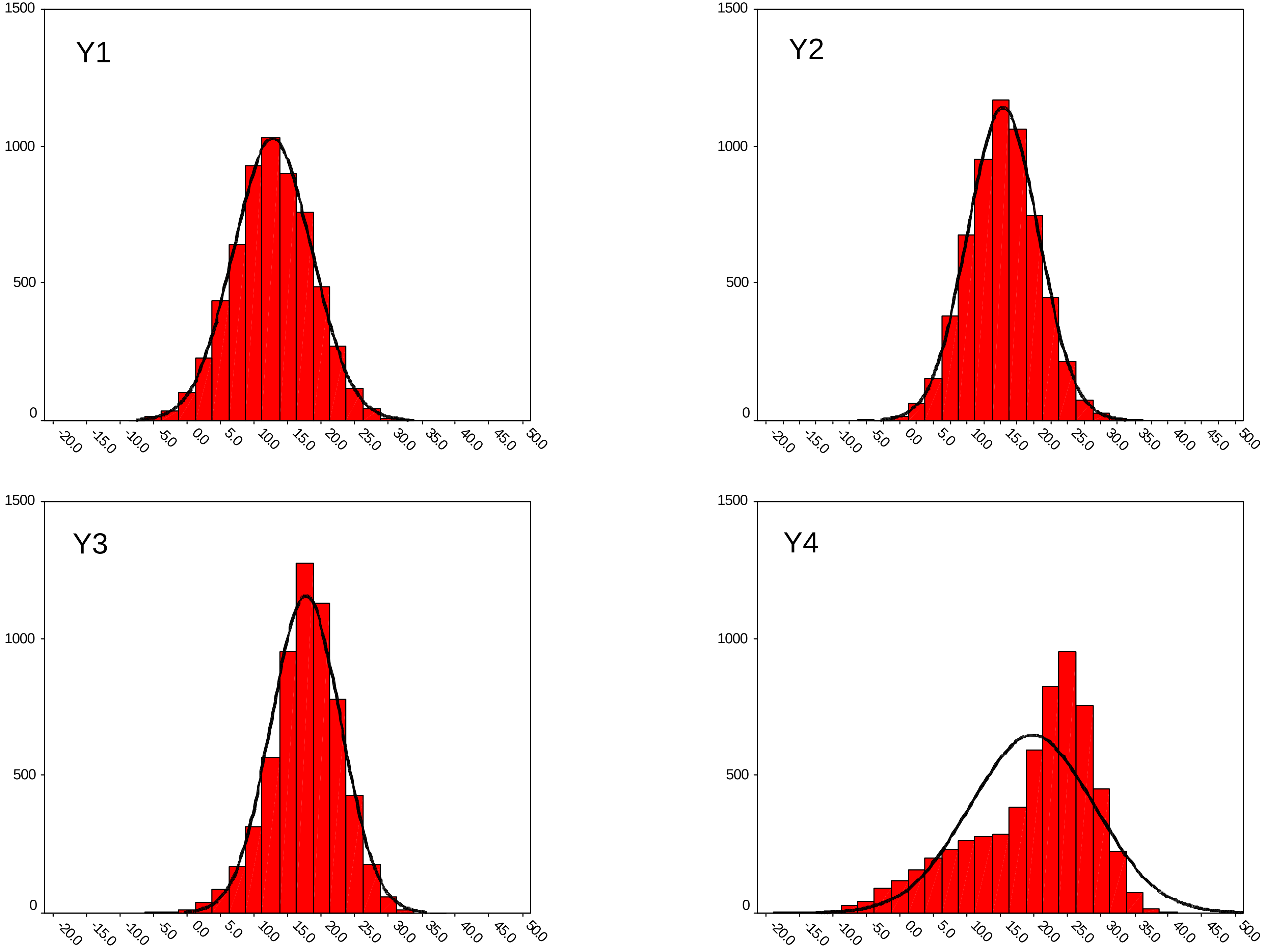

Figure 3. Distributions of Growth Parameters For MultiClass Data

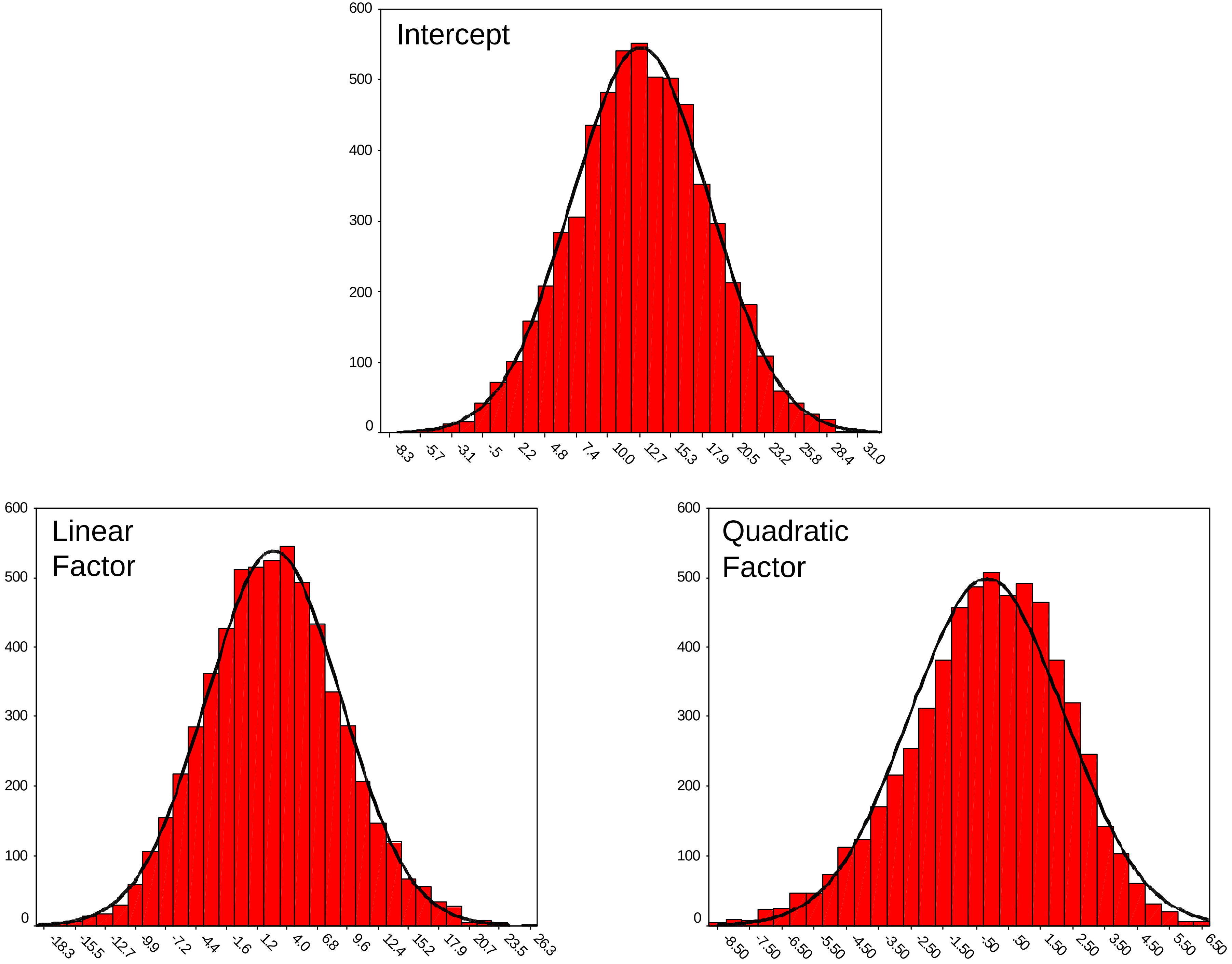

Figure 4. Fit Statistics for GMM Models

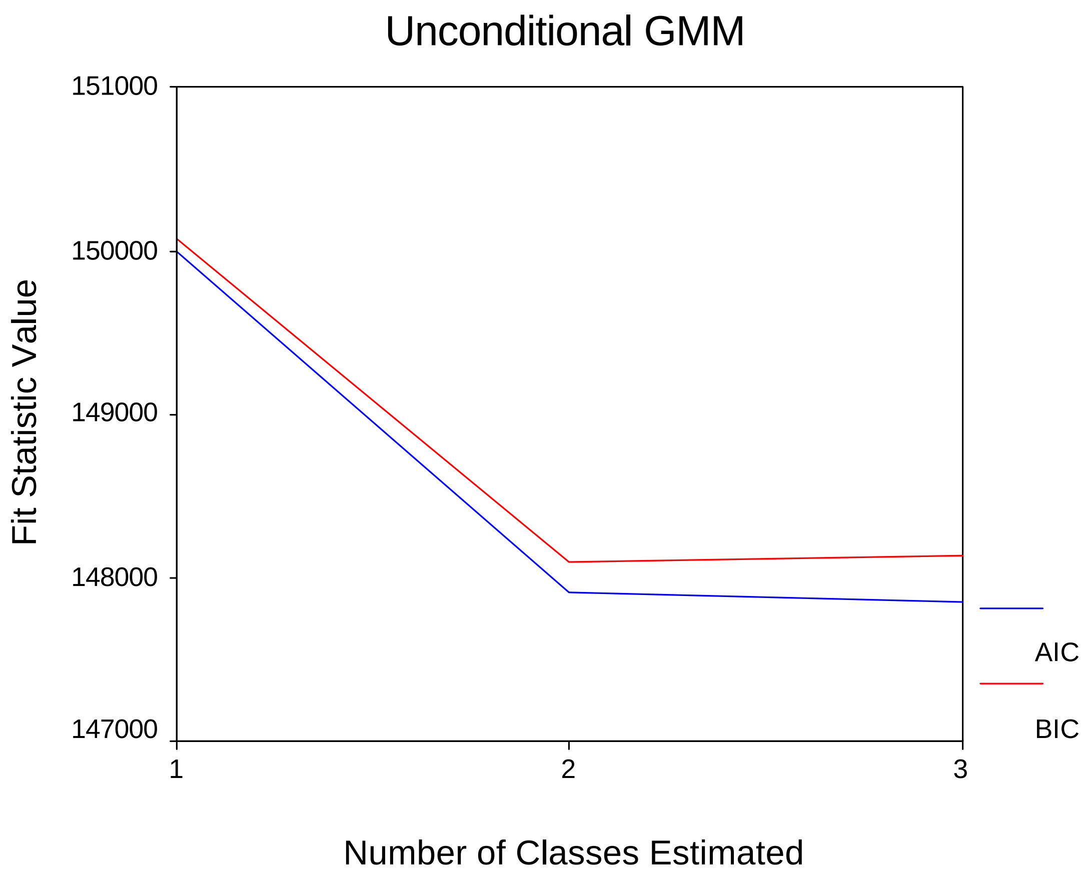

Figure 5. Misspecified Model Implied Trajectories for Multiclass Data