Comparing the evidence for categorical versus dimensional representations of psychiatric disorders in the presence of noisy observations: a Monte Carlo study of the Bayesian Information Criterion and Akaike Information Criterion in latent variable models.

Michael Hallquist,

Thomas M. Olino, &

Paul A. Pilkonis

University of Pittsburgh
Introduction

- Considerable debate has centered on whether psychopathology is best represented by continuous dimensions or categorical subtypes (First, 2005; Kramer, Noda, & O’Hara, 2004).

- Prior statistical research has focused on the ability of model selection criteria (e.g., the Bayesian Information Criterion) to adjudicate the latent structure of psychometric indicators.
Introduction

Information-Theoretic Latent Distribution Modeling: Distinguishing Discrete and Continuous Latent Variable Models

Kristian E. Markon and Robert F. Krueger
University of Minnesota, Twin Cities Campus

- Markon & Krueger (2006): selection criteria (esp. BIC) can be used to probe the loss of statistical information resulting from discretizing an underlying dimension.
  - e.g., if latent structure is dimensional, estimating a few points along a latent distribution (oligovalued) may be less informative than a parametric distribution approximating the dimension (e.g., normal).
Introduction

- Factor mixture models that combine categorical and dimensional structure can often be resolved by model selection criteria (Lubke & Neale, 2005).
- Resolution of dimensional, categorical, and hybrid latent structure is clearest when
  1. indicators are well separated across latent classes
  2. there are more indicators
  3. there is little unmodeled within-class residual association among indicators
Introduction

- Yet simulation studies have only tested multivariate normal (or skew-normal) data with little attention to the influence of scatter/outliers on latent structure decisions.
- There has been recent interest in the robustness of clustering methods to data that include scatter (Maitra & Ramler, 2008).
- In the case of empirical studies of mental disorders, data are rarely normal and heterogeneity sometimes manifests as indicators that are difficult to cluster.
Scientific question

- If mental disorders in the population were sorted by severity (e.g., low, medium, high), would we be able to resolve categorical versus dimensional structure when some observations are noisy?

- Parameterization: latent profile model (LPM) with diagonal covariance matrices.

\[
f(y) = \sum_{k=1}^{K} P(k) f(y|\mu_k, \Sigma_k).
\]

\[
y_k \sim N(\mu_k, \Sigma_k).
\]
Model selection criteria

- Generally: \[-2\ell(\theta | x) + \alpha \kappa.\]

- Akaike Information Criterion (AIC):
  - \(\alpha = 2\)

- Bayesian Information Criterion (BIC):
  - \(\alpha = \log(n)\)

- Finite-sample adjusted AIC (AIC\(_C\))
  - \(\alpha = 2\cdot [n/(n-\kappa-1)]\)
Asymptotic properties of AIC and BIC

- **AIC**
  - Efficient: tends to select the model that minimizes prediction error as sample size increases.
  - Primarily concerned with minimizing the relative Kullback-Leibler (K-L) distance between a statistical model and the unknown mechanisms giving rise to the observed data.

- **BIC**
  - Consistent: “true model” will be selected with increasing probability as sample size increases.

Vrieze 2012; Yang 2005
Simulation Conditions

- Number of latent classes: 3 (equal prob.)
- Number of indicators: 5
- Sample size: 45, 90, 180, 360, 720, 1440
- Mean separation: .25, .5, 1.0, 1.5, 2.0, 3.0
- Proportion of noisy observations added:
  - 0%, 2.5%, 5%, 10%, 15%, 20%, 25%, 30%
- Number of replications: 200/condition
Simulation methods

- Within-class multivariate normal data generated using the MixSim package in R.
- Noise observations added by randomly sampling observations that were outside the 99% ellipsoids of concentration for all clusters (Maitra & Ramler, 2009).
- Noise observations were uniformly distributed with bounds $-1.5 \times \text{IQR} < x < 1.5 \times \text{IQR}$
Simulation Pipeline

For each replication, simulate mvnorm Gaussian mixtures using MixSim in R

Add some proportion (0–0.3) of scatter observations outside of 99% ellipsoid bounds

Analyze simulated data in Mplus 7.0 using LPM and CFA

Summarize and visualize results across conditions
Example of clean latent structure

\( n = 720, \text{ LC separation } = 2.0 \text{ SD}, 0\% \text{ Noise} \)

![Histograms showing latent class structure](image-url)
Example of noisy latent structure

\( n = 720, \) LC separation = 2.0 SD, 10% Noise
When indicators were separated by 1SD or less, scatter observations (esp. > 10%) seriously degraded latent structure.

For models where indicators were well separated (2–3 SDs), parameter coverage for LPMs was relatively robust to scatter.

Adding scatter to data separated by < 1SD shifted evidence toward categorical model.

For indicators separated by > 1SD, adding scatter weakened evidence for categorical model.

Greater latent separation was more important than sample size in determining coverage of latent means.
Discussion

- For moderate sample sizes ($n = 360$ or less) and low latent separation (0.75SD or less) BIC tends to select dimensional model, whereas AIC prefers categorical model regardless of scatter.
- $AIC_C$ fell in the middle, but has a steeper penalty than BIC at small sample sizes.
Future directions

- Exploring how scatter observations are clustered (normalized mutual information).
- New approaches to scatter: corruption of selected observations by adding random deviate.
- Testing clustering algorithms (e.g., k-clips; Maitral & Ramler, 2009) that identify scatter and remove these observations from substantive clusters.
- Testing performance of model selection criteria for dimensional data with scatter.