Bayesian Mixture Modeling

Sarah Depaoli
University of California, Merced

July 21, 2014
Mplus Users Meeting, Utrecht

Organization of the Talk

Bayesian Mixture Modeling

Sarah Depaoli
UC Merced

Intro
Organization
Mixtures
Bayesian

Example

Simulations

Benefits

Cautions

Conclusions

- Mixture modeling
- Bayesian estimation framework
- Motivating examples
  - Introduce the basic LCA model
    - Illustrated using Adult California Tobacco Survey
  - Priors for LCA
  - Bayesian portions of M*plus* code
    - Illustrated using Youth Risk Behavior Survey
- Simulation findings: Estimating mixture models in M*plus*
- Benefits of Bayes for mixture models
- Cautions using Bayes with mixture models
- Concluding remarks

Introduction to Mixture Modeling

- A large part of so-called second-generation SEM is the ability to model different *unobserved* groups of individuals.

- These unobserved groups can be captured through mixture models, and substantive differences across the groups can be identified.

- Mixture modeling has proved to be a useful tool for accounting for heterogeneity within a population, and the flexibility of mixture models has allowed for some innovative modeling techniques.[1]

[1] For detailed information about mixture modeling, see: McLachlan, G., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Introduction to Mixture Modeling cont.

- Mixtures are particularly useful when there is unobserved heterogeneity in the data.
- In this case, the variables that cause heterogeneity in the data are not known prior to data analysis.
- The researcher hypothesizes that the population is comprised of an unknown (precise) number of substantively distinct subpopulations (or classes).
  - e.g., depression (subpopulations: non, mild, depressed); reading achievement (subpopulations: low, average, high)
- Typically, it is "known" that the population is heterogeneous, but there is no known indicator of class membership.
- Class membership must be determined based on observed data patterns.

Bayesian Estimation Framework

- The Bayesian estimation framework is a complex system used to estimate models.[2]
- One of the key differences between frequentist (e.g., ML/EM) and Bayesian estimation is the use of *prior distributions*.

$$Posterior = Data * Prior \tag{1}$$

- The prior distribution is moderated by the data and this relationship produces the *posterior distribution* (estimate).

[2] For more details about Bayesian estimation, see: Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

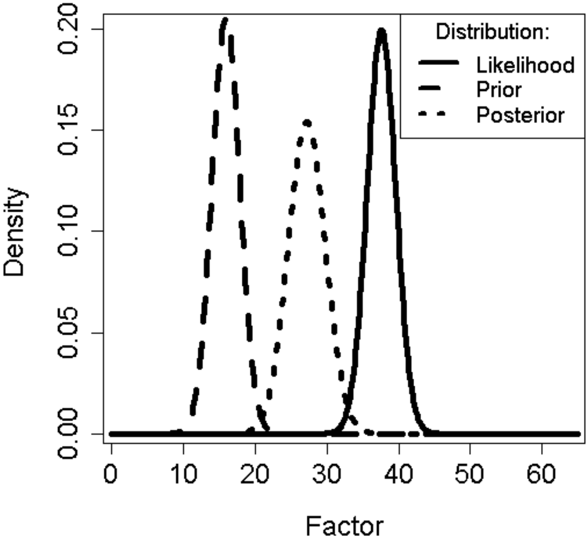
Bayesian Estimation cont.

- Prior distributions are placed on every model parameter that we wish to estimate.
  - Think of priors similar to a "bet".
- These distributions represent the amount of uncertainty that we have surrounding the parameters in our model.
- Specifically, priors represent our opinions about each model parameter.

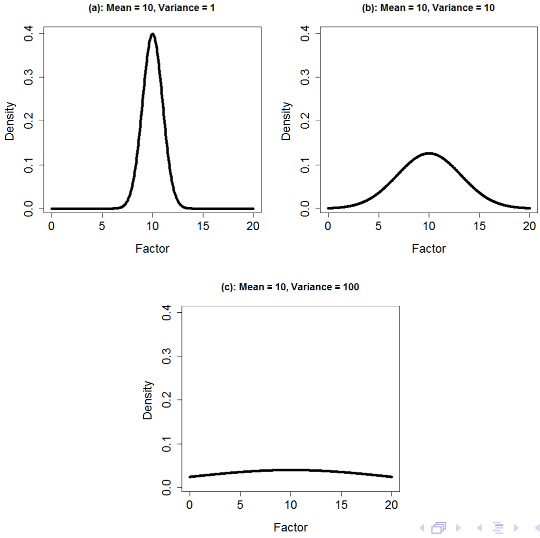
Bayesian Estimation cont.

- Priors allow us to incorporate our (un)certainty about model parameters using probability distributions.
  - For example, Intercept $\sim \mathcal{N}(\mu, \sigma^2)$.
    - The μ and σ^2 terms are called *hyperparameters*.
- In addition, we can also make an assumption about the particular values that the intercept can take on.
  - Diffuse: having no idea about the parameter value.
  - Informative: having a very strong idea about the parameter value.
  - Weak: Using less information then is available for the prior.
- The specification of these prior distributions is an integral part of using the Bayesian estimation framework.

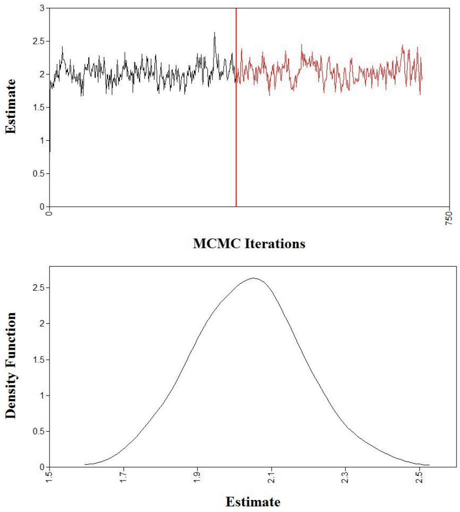
Contrived Example of the Bayesian Framework



Priors: Different Levels of Informativeness



Posterior Chain and Corresponding Density



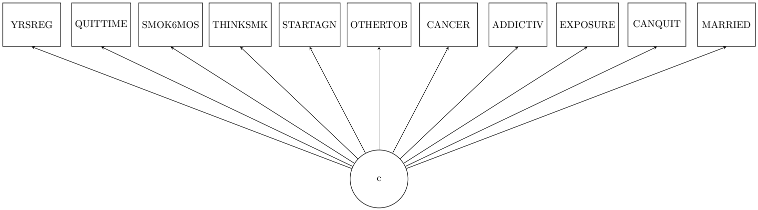
Motivating Example: LCA

- Latent class analysis (LCA) is a method used to capture different response patterns for discrete observed variables.
- With these types of variables, there can be a very large number of response patterns (e.g., 5 binary items: $2^5 = 32$ response patterns).
- LCA summarizes these response patterns into a few, substantively meaningful latent classes.

Latent Class Analysis cont.

- Each individual has a probability for belonging to each class, and the individual is assigned to the class corresponding with the highest probability of membership.
- The assigned class consists of individuals with similar response patterns.
- LCA provides a succinct description of the latent classes through the different patterns of responses on the observed variables.

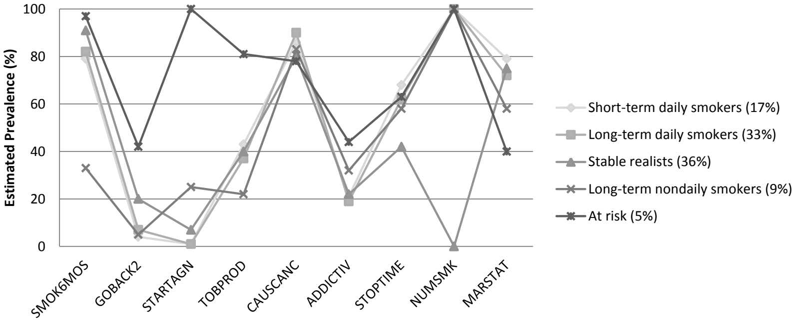
LCA: Example of Former Smokers[3]



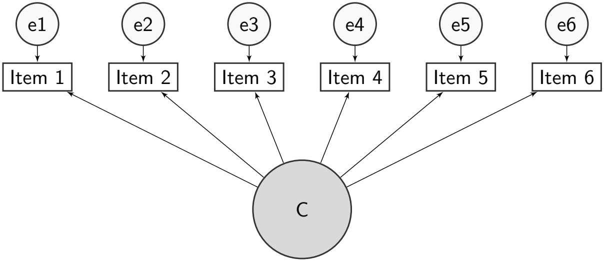
- Using ∼2500 cases from the Adult California Tobacco Survey to identify possible latent groups of former smokers

[3] Clifton, J., Depaoli, S., and Song, A. (under revision). Are all former smokers alike? A latent class analysis.

LCA: Example of Former Smokers cont.



Priors for a Basic LCA



- Response probabilities: Normal prior, N(0,10^{10})
- Latent class proportions: Dirichlet prior, D(10,10)
- Residual variances: Inverse-Gamma prior, IG(-1,0)

General M*plus* code for Bayesian Estimation

```
ANALYSIS:
  TYPE=MIXTURE;
  ESTIMATOR = BAYES;
  CHAINS=1;
  DISTRIBUTION=50,000;
  POINT=MODE;
  ALGORITHM = GIBBS (PX1);[4]
  BCONVERGENCE=.05
  BITERATIONS=50,000 0;
  FBITERATIONS = 50,000;
  THIN=1;
```

---

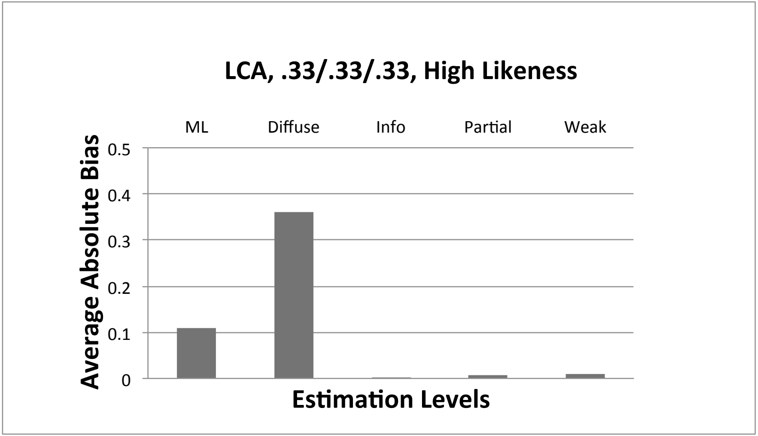
[4] More information about Bayesian samplers in: Asparouhov, T, and Muthén, B. (2010). Bayesian analysis using M*plus*: Technical implementation. Technical Report. Los Angeles: Muthén & Muthén.

M*plus* code for Bayesian LCA (Collins & Lanza, Youth Risk Behavior 2007, Priors from 2005)


```
model priors:
d1~D(8917,720);
d2~D(1211,720);
    !c#1
j11~N(-3.132,0.084);
j12~N(-3.892,0.112);
j13~N(-5.141,0.317);
    !c#2
j21~N(1.150,0.174);
j22~N(-0.770,0.121);
j23~N(-1.746,0.158);
    !c#3
j31~N(0.611,0.134);
j32~N(0.674,0.098);
j33~N(-0.163,0.093);
```

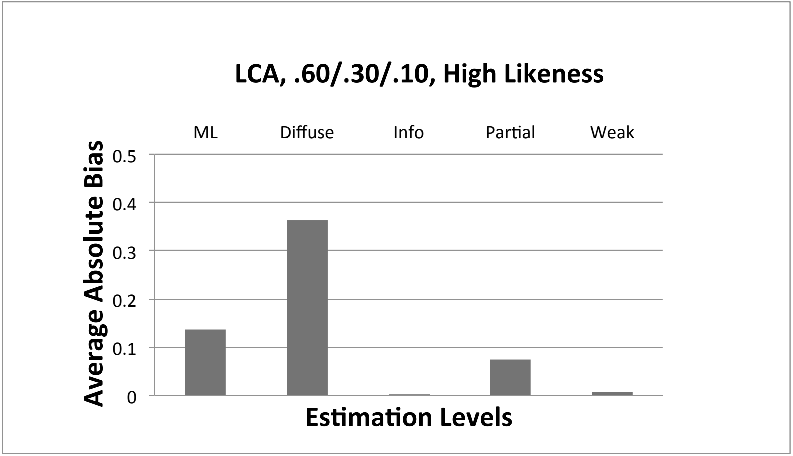
```
  %overall%
[c#1*2.595](d1);
[c#2*0.587](d2);
    %c#1%
[cig13$1*-3.132](j11);
[cig30$1*-3.892](j12);
[drive$1*-5.141](j13);
    %c#2%
[cig13$1*1.150](j21);
[cig30$1*-0.770](j22);
[drive$1*-1.746](j23);
    %c#3%
[cig13$1*0.611](j31);
[cig30$1*0.674](j32);
[drive$1*-0.163](j33);
```

LCA Simulation Results[5]



[5]Depaoli, S. and Clifton, J. (in preparation). The specification and impact of prior distributions for categorical latent variable models.

LCA Simulation Results cont.



Benefits: General

- In general, the Bayesian approach can produce a drastic improvement in accuracy in parameter estimates and mixture class proportions, especially when more informative (or weak) priors are specified.

- This approach can also help identify small but substantively real latent classes.[6]

[6] see e.g., Depaoli (2013); and van de Schoot, Depaoli, van Loey, N., and Sijbrandij, M. (under review). Integrating background knowledge about traumatic stress experienced after trauma into latent growth mixture models.

Benefits: Accurate, Informative Priors

- Informative priors perform quite well in simulation in that they are largely able to uncover small but substantively different mixture classes.[7]
- Latent class proportions are also well recovered with the use of (weakly) informative Dirichlet priors on the class proportions.[8]

[7] Depaoli, S. (2013).
[8] Depaoli, S. and Clifton, J. (in preparation).

Benefits: Inaccurate Priors

- Mixture modeling is relatively robust to inaccuracies in prior distributions.[9]
- This is an important finding given that if the location of a prior distribution is very wrong, then the parameter value can still be accurately recovered by even moderately increasing the variance hyperparameter of the prior.
- One area not examined yet is the inaccuracy of the Dirichlet prior and the impact this would have on substantive findings.

[9] Depaoli, S. (2014).

Cautions: Diffuse Priors

- Findings have suggested that more informative priors are necessary in the context of mixture modeling.[10]
- Diffuse priors (e.g., $N(0, 10^{10})$) may have a more harmful impact on parameter estimates than inaccurate priors in the case of mixture modeling.

[10] Depaoli (2012); Depaoli (2013)

Concluding Remarks about Bayesian Mixture Modeling in M*plus*

- The Bayesian estimation framework shows promise in accurately estimating mixture models under various modeling conditions.
- Recognizing that our priors will undoubtedly contain some level of inaccuracy according to the unknown population, it is important to conduct a sensitivity analysis in order to assess how much of an impact different levels of the prior have on model results.
- Openness and transparency are vital for implementing any statistical tool, but this is especially the case for Bayesian tools.

Thank You!

Questions or Comments:

Sarah Depaoli: sdepaoli@ucmerced.edu

Specification of Latent Class Analysis

- The probability of membership in latent class c is represented by γ_c and

$$\sum_{c=1}^{C} \gamma_c = 1. \tag{2}$$

- For a given observed item j, the probability of response r_j given membership in class c is given by an item-response probability $\rho_{j,r_j|c}$. Note that the vector of item-response probabilities for item j conditional on latent class c always sums to 1.0 across all possible responses to item j as denoted by

$$\sum_{r_j=1}^{R_j} \rho_{j,r_j|c} = 1 \tag{3}$$

for all j observed items.

Specification of Latent Class Analysis

- In order to define the LCA model, the probability of a given pattern of responses must be computed. Let y_j represent the j element for the observed response pattern denoted as vector $\mathbf{y}$. Next, let $I(y_j = r_j)$ represents in indicator variable such that the indicator variable equals 1 when variable $j = r_j$ and 0 otherwise. Then, the probability of observing a particular set of item responses $\mathbf{y}$ can be written as

$$P(\mathbf{Y} = \mathbf{y}) = \sum_{c=1}^{C} \gamma_c \Pi_{j=1}^{J} \Pi_{r_j=1}^{R_j} \rho_{j,r_j|c}^{I(y_j=r_j)}. \quad (4)$$

Specification of Latent Class Analysis

- Essentially, Equation 4 indicates that the probability of observing a particular response pattern $\mathbf{y}$ is a function of the probability of membership in each of the C latent classes given by the γ_c term and the probability of each response conditional on latent class membership denoted by $\rho_{j,r_j|c}$. To provide an example of more concrete notation, Equation 4 can be expanded out for observed categorical items $j_1, \ldots, j_4$ respectively:

$$P_{j=1,\ldots,4} = \sum_{c=1}^{C} \gamma_c \rho_{j=1|c} \rho_{j=2|c} \rho_{j=3|c} \rho_{j=4|c}, \tag{5}$$

- where the probability of a given response pattern for items $j = 1, \ldots, 4$ is a product of the proportion of individuals in latent class c and response probabilities for observed items $j = 1, \ldots, 4$ conditioned on class membership.

Bayesian Estimation

- We first set up the joint probability distribution which is

$$P(\mathbf{y}, \mathbf{\Theta}) = P(\mathbf{y}|\mathbf{\Theta})P(\mathbf{\Theta}). \tag{6}$$

- Bayes theorem uses this product to compute the probability of $\mathbf{\Theta}$ given the observed data $\mathbf{y}$ through the following

$$P(\mathbf{\Theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{\Theta})P(\mathbf{\Theta})}{\int_{\mathbf{\Theta}} P(\mathbf{y}|\mathbf{\Theta})P(\mathbf{\Theta})d\mathbf{\Theta}} \tag{7}$$

- where $P(\mathbf{y}|\mathbf{\Theta})$ represents the likelihood (the observed data given the distributional parameters), $P(\mathbf{\Theta})$ represents something called a prior distribution that is coupled with the likelihood, and $P(\mathbf{\Theta}|\mathbf{y})$ is the posterior distribution of $\mathbf{\Theta}$.

Inverse-Wishart Specification in M*plus*

- There are three specifications of the inverse Wishart that are discussed as non-informative in Asparouhov and Muthén (2010)[12]
- The first specification is IW(0,-p-1), which is the current default setting in M*plus* version 7.2 for covariance matrices and mimics a uniform prior bounded at $(-\infty,\infty)$.
- The second specification is IW(0,0).
- The last specification discussed is IW(I,p+1), where this prior mimics the case where off-diagonal elements (covariances) of the covariance matrix would have uniform priors bounded at [-1,1] and diagonal elements (variances or residual variances) distributed as IG(1,.5).

[12] page 35: Bayesian analysis using M*plus*: Technical implementation. Technical Report. Version 3.

Cautions: Inverse Wishart Priors

- Although not discussed here, it is important to address some points about the inverse Wishart prior, which is commonly implemented with mixture models (for covariances).
- Changing default M*plus* settings of this prior may create a multivariate prior that is non-positive definite.
  - When the default is changed, you have univariate inverse gamma priors on diagonals and univariate uniform or normal priors on off-diagonals.[13]
- When in doubt, seek advice from a statistician!

[13] For more details see: Depaoli and van de Schoot (under review).

References

- Asparouhov, T, and Muthén, B. (2010). Bayesian analysis using M*plus*: Technical implementation. Technical Report. Los Angeles: Muthén & Muthén.
- Clifton, J., Depaoli, S., and Song, A. (under revision). Are all former smokers alike? A latent class analysis.
- Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*, 239-252.
- Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186–219.

References cont.

- Depaoli, S. (2012). Measurement and structural model class separation in mixture-CFA: ML/EM versus MCMC. *Structural Equation Modeling: A Multidisciplinary Journal*, *19*, 178-203.
- Depaoli, S. and Clifton, J. (in preparation). The specification and impact of prior distributions for categorical latent variable models.
- Depaoli, S. and van de Schoot, R. (under review). The WAMBS-checklist: When to worry, and how to avoid the misuse of Bayesian statistics.
- Diebolt, J., and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society*. Series B (Methodological), 363-375.

References cont.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.
- McLachlan, G., and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.
- Natarajan, R., and McCulloch, C. E. (1998). Gibbs sampling with diffuse proper priors: A valid approach to data-driven inference?. *Journal of Computational and Graphical Statistics, 7*(3), 267-277.
- Richardson, S., and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology), 59*(4), 731-792.

References cont.

- Roeder, K., and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association, 92*(439), 894-902.
- van de Schoot, Depaoli, van Loey, N., and Sijbrandij, M. (under review). Integrating background knowledge about traumatic stress experienced after trauma into latent growth mixture models.