Published in final edited form as:

Commun Stat Simul Comput. 2018 ; 47(4): 1028–1038. doi:10.1080/03610918.2017.1303726.

An evaluation of the bootstrap for model validation in mixture models

Thomas Jaki,
Lancaster University

Ting-Li Su,
Manchester University

Minjung Kim, and
The University of Alabama

M. Lee Van Horn
University of New Mexico

Abstract

Bootstrapping has been used as a diagnostic tool for validating model results for a wide array of statistical models. Here we evaluate the use of the non-parametric bootstrap for model validation in mixture models. We show that the bootstrap is problematic for validating the results of class enumeration and demonstrating the stability of parameter estimates in both finite mixture and regression mixture models. In only 44% of simulations did bootstrapping detect the correct number of classes in at least 90% of the bootstrap samples for a finite mixture model without any model violations. For regression mixture models and cases with violated model assumptions, the performance was even worse. Consequently, we cannot recommend the non-parametric bootstrap for validating mixture models.

The cause of the problem is that when resampling is used influential individual observations have a high likelihood of being sampled many times. The presence of multiple replications of even moderately extreme observations is shown to lead to additional latent classes being extracted. To verify that these replications cause the problems we show that leave-k-out cross-validation where sub-samples taken without replacement does not suffer from the same problem.

Keywords

Finite mixture models; leave-k-out cross-validation; model validation; nonparametric Bootstrap; regression mixture models

Introduction

Finite mixture models and their extensions – regression mixture models – are widely used for modeling population heterogeneity. These models have been used in diverse areas

including modeling heterogeneity in patterns of substance use (M. L. Van Horn et al., 2008), allowing differences between countries in item measurement properties (de Jong & Steenkamp, 2010), and assessing heterogeneity in the effects of a predictor on an outcome (Lenk & DeSarbo, 2000). The estimation of mixture models relies on strong assumptions such as conditional independence or a specific shape (typically normal) for the residual distribution (Bauer & Curran, 2003; M. L. Van Horn et al., 2012). Because of the sensitivity of these models to their assumptions, methods for model validation and robust approaches for estimating sampling distributions are crucial.

One commonly used approach for model validation and estimating complex sampling distributions is the bootstrap (Davison & Hinkley, 1997; Efron & Tibshirani, 1993). Bootstrap methods have been used both for providing empirically derived sampling distributions and specifically to obtain standard errors (Basford, Greenway, McLachlan, & Peel, 1997; McLachlan & Peel, 2000; Newton & Raftery, 1994), which are robust to distributional assumptions (DiCiccio & Efron, 1996). The other use of bootstrap methods is as diagnostic tools for validating mixture model results (Grün & Leisch, 2006, 2010). Although methods for validating mixture models are clearly needed and the bootstrap being commonly used for this purpose, our experience suggests that the use of the non-parametric bootstrap with mixture models is problematic. This appears to be because individual data points can be replicated several times which will result in a high chance that influential points are over-represented resulting in additional classes being identified. In this paper we use probability arguments and simulations to show that the non-parametric bootstrap is not suitable for validating finite mixtures and regression mixture models. This is despite the fact that this method has been suggested previously (Grün & Leisch, 2004; Schlattmann, 2005; M. L. Van Horn et al., 2009) and is for example implemented in the widely used FLEXMIX package (Grün & Leisch, 2008; Leisch, 2004). To further illustrate that the problems arise from replicated observations, we consider the leave-k-out cross-validation (also called leave-k out jackknife; Davison & Hinkley, 1997) which does not suffer from replicated observations in the same way as the bootstrap and appears to work more satisfactorily provided that the sample size used is sufficiently large.

Motivational example

To illustrate the problem which non-parametric bootstraps can create, Figure 1 displays the histogram of 100 randomly drawn observations from a standard normal distribution together with the histogram of one non-parametric bootstrap resample, i.e. observations are randomly drawn with replacement from the original data. The original data display a single value that is at around 3 relatively far above the other observations. In the resampled dataset this value is randomly selected 3 times. Similarly, on the left side of the distribution the seven smallest observations in the original data were replicated a total of ten times. The potential consequence of these artificial “clusters” is that additional classes are introduced by using the bootstrap.

In this particular example fitting one and two-class finite mixture models to the data using MPLUS 7.0 (Muthén & Muthén, 2010) and using the BIC to choose which model is the most appropriate selects a single class model for the original data. Using the same procedure

for the resampled data, results in selection of a 2-class model. In this case, the process of resampling has artificially introduced an additional class despite the original data satisfying all model assumptions perfectly. What is quite surprising, however, is that the additional class does not appear to capture the extreme observation to the right, but instead represents the only somewhat exaggerated peak to the left of the distribution. This suggests that not only over-replicating single extreme values can introduce new classes, but a moderate exaggeration of several values can already yield additional classes.

One immediate question arising from this example is, whether this example is just a fluke or if over-representation of values, particularly in the tail of the distribution, is common. To study this, we consider a sample of n subjects and begin by considering the probability of a particular observation, such as the largest observation in the example, being replicated at least q times in a bootstrap sample of size n . This probability can be found as

$$P(X \geq q) = 1 - \sum_{l=0}^{q-1} \binom{n}{l} \left(\frac{1}{n}\right)^l \left(\frac{n-1}{n}\right)^{n-l}$$

where X is the number of times the observation of interest is selected. Based on this, one can find that the probability of replicating one value at least 3 times (as with the largest value in the example) is almost 8% when $n=100$. In the example it appears, however, that a small set of observations that have been over-represented resulted in an extra class. Denoting Y as the total number of times m observations are selected, we can find the probability of selecting these at least q times as

$$P(Y \geq q) = 1 - \sum_{l=0}^{q-1} \binom{n}{l} \left(\frac{m}{n}\right)^l \left(\frac{n-m}{n}\right)^{n-l}$$

The probability of selecting any of the 7 smallest observations at least 10 times in a sample of size 100, as in the example, is then about 16%.

The above provides some insights into the potential problems non-parametric bootstrap could encounter when used to verify the results of mixture models. Across many bootstrap samples the probability that a given influential case is replicated multiple times becomes quite high, and the probability that a set of influential cases is over represented is even higher. It is, however, clear that just because a single value is replicated several times does not necessarily lead to an additional class in any one resampled dataset. When oversampling of influential data points becomes a problem in mixture models is not clear. In this paper, we use simulations to examine the use of the non-parametric bootstrap to verify model solutions of finite mixture and regression mixture models. We will focus on using the bootstrap (and subsequently also leave-k-out cross-validation) to verify class enumeration and the ability of the bootstrap to provide robust confidence intervals for the model parameters. In the next sections we will briefly review the regression mixture models and the bootstrap resampling methods for model validation.

Regression mixture models.

Regression mixture models are a generalization of the finite mixture model in which the density of an outcome variable y is modeled as a weighted sum of K component distributions,

$$f(y|\phi, X) = \sum_{k=1}^K \pi_k f_k(y_k|\theta_k, X) \quad \text{eq. 1}$$

where ϕ denotes all parameters to be estimated. In a regression mixture model the density in each class is conditional on a set of regression weights which may be specific to that class, resulting in a within class model for the outcome:

$$y_i|k, X_i = \beta_{0k} + \beta_{1k} X_i + \varepsilon_{ik} \quad \text{eq. 2}$$

where y is the outcome variable which has a class specific intercept β_{0k} (which simplifies to the mean when there are no covariates), which is modeled as a function of a matrix of covariates, X_i , whose class specific regression weights are contained in the vector β_{1k} . The random errors are assumed to be class specific, $\varepsilon_{ik} \sim N(0, \sigma_k^2)$. The number of components or classes, K , is specified, but the class prevalence in the population, $\pi_1, \dots, \pi_K$, are estimated. Regression mixtures are differentiated from the broader family of mixture models by the presence of class specific regression weights which allow the effects of covariates to be uniquely estimated in each class.

In practice there are two parts to the estimation of a mixture model: 1) latent class enumeration in which the number of classes to be used in the model is determined, and 2) estimation of the parameters that define each class. While latent class enumeration may be driven by theoretical hypotheses in some instances, it is more often determined by estimating a series of models with increasing numbers of classes and then choosing the number of classes to be carried forward based on model fit to the data. The models are then typically compared using penalized information criterion. The Bayesian Information Criteria (BIC) and sample-size adjusted BIC (aBIC) will be the primary focus of this paper because it has consistently demonstrated good performance for model selection in regression mixtures (George et al., 2013; M. L. Van Horn et al., 2012).

An important feature of all mixture models are the assumptions made about error terms which are required to identify model parameters (McLachlan & Peel, 2000). When working with continuous outcome variables an assumption of normality is common, although it may be replaced by a skew normal (Liu & Lin, 2014; M. L. Van Horn et al., 2012) or Poisson (Lanza, Kugler, & Mathur, 2011) distribution in some cases. When working with categorical or ordinal outcomes, assumptions such as local independence (Bauer & Curran, 2004) or proportional odds (George et al., 2013) may also be used. Parameter estimates have been

shown to be highly sensitive to violations of the identifying assumptions made for any particular mixture model (M. L. Van Horn et al., 2012). The sensitivity of these models suggests the need for more robust approaches to both evaluate model stability and to assess sampling distributions for model parameters.

Bootstrapping.

Methods using resampling are well established approaches to empirically estimate sampling distributions without relying on particular distributional assumptions. Perhaps the most well-known of these methods is the non-parametric bootstrap which involves taking repeated samples, with replacement, from a dataset. Variability between these samples mimics the variability seen when taking samples from a population and can thus be used to derive the sampling distribution of parameters by estimating the original model separately in each bootstrapped sample and then examining the resulting distribution of the parameter of interest across all bootstrapped samples. This is called the non-parametric bootstrap because it involves resampling observed data rather than sampling from a particular parametric distribution (Efron & Tibshirani, 1993). The parametric bootstrap, in contrast, involves sampling from the model implied data given that the model is true and all model assumptions are met. Clearly, if the rationale for using the bootstrap is based on the sensitivity of the model to assumptions, then the parametric bootstrap is not useful as it assumes that all model assumptions are met. As the objective of this paper is to evaluate ideas for model validation, it focuses on the non-parametric bootstrap.

While initially developed as a method for providing empirical estimates of sampling distributions, because it mimics random draws from a population, the non-parametric bootstrap has also been proposed as a method for validating model results (Thompson, 1993, 1995). This is especially relevant with finite mixture results where there is a concern that small, possibly chance, deviations from assumed distributions and multivariate outliers could greatly impact model results. In this case, the bootstrap is intended to assess the replicability of a particular set of results in random draws from the population (M. L. Van Horn et al., 2009) as well as providing a diagnostic tool for assessing model sensitivity (Grün & Leisch, 2004). Based on limited simulation conditions, the use of the non-parametric bootstrap method has been specifically recommended for finding the optimal number of latent classes in finite mixture models (Schlattmann, 2003, 2005).

The simulations in this paper evaluate the use of non-parametric bootstrapping in validation results of finite and regression mixture models for both latent class enumeration and parameter estimation. If our previously developed hypothesis about the effects of resampling with replacement on mixture models is correct, we expect that bootstrapping will lead to selecting models with too many latent classes and to less stable model solutions. We test this hypothesis by also evaluating leave- k -out cross-validation, which uses resamples of size $n-k$ without replacement. We hypothesize that under ideal conditions cross-validation will not exhibit the same problems as the non-parametric bootstrap, confirming that sampling with replacement causes problems.

Methods

To systematically evaluate if non-parametric bootstrap can be used for model validation we utilize Monte-Carlo simulations with 100 independent datasets of size n from two different models. The first model, a finite mixture model, assumes that each of two groups is normally distributed with equal variance of 1 but means of 0 and 2. More specifically, observations from group k are generated as

$$y_{ik} = \mu_k + \varepsilon_i$$

where $\varepsilon \sim N(0,1)$, $\mu_1=0$ and $\mu_2=2$. The second model, a 2-class regression mixture model, considers a linear regression for each subgroup with different slope and intercept;

$$y_i | k, X_i = \beta_{0k} + \beta_{1k} X_i + \varepsilon_{ik}$$

where $\varepsilon_{ik} \sim N(0, 1 - \beta_{1k}^2)$ and the covariate X is standard normally distributed. The intercepts, β_{0k} , are 0 and 0.5 and the slopes, β_{1k} , are 0.2 and 0.7. The variance of the error terms, ε_{ik} , are chosen so that the variance of y_k will be 1 in each class. This specification allows the slopes to be interpreted as the correlation between x and y and are chosen to represent a strong and a weak correlation. For both models we assume that an equal proportion of subjects is in each class (i.e. $\pi_k=0.5$) and simulate a sample size of 3000 in each class. For each of the independent samples we then generate 100 bootstrap resamples of size n by sampling with replacement from the independent datasets.

In addition to simulating data satisfying the assumptions of mixture models, we also evaluate the situation, where p additional outliers contaminate the data. For the finite mixture model $p=5$ outliers are generated from a normal distribution with mean of -4 and variance 0.01 while for the regression mixture model $p=10$ or 30 outliers are generated from a normal distribution with mean of 2.5 and no effect of the covariate. Figure 2 displays the distribution/scatter plot for the data generation scenarios with outliers. It is worth pointing out, that even the 30 outliers make up less than 0.5% of the total dataset and are visually undetectable in these graphs. Consequently, one would naively expect that they will not influence the results of these models and that, even if they are influencing model fit, any method for model validation or robust estimation method should be able to cope with them.

In the first evaluation we compare the ability of standard model fit criteria to identify the correct number of classes in independent samples to the resampled data. We fit one, two and three-class models to each of the independent datasets as well as to 100 bootstrap resamples of each of the independent datasets and evaluate how frequently the correct 2-class solution is found. We use the BIC and aBIC to choose the correct number of classes.

The second evaluation assumes that the correct two-class model is found and examines the distribution of each model parameter across bootstrap draws. If non-parametric bootstrap can be utilized to estimate the sampling distribution of the model parameters, given that the correct number of classes are somehow identified, then we would expect that the coverage

observed (the percentage of simulations for which the true value for each parameter was inside the 95% bootstrap confidence interval) to be equal to the confidence interval chosen (95%). To do so, we use the data generation structure as in the first evaluation; i.e. 100 independent datasets with 100 bootstrap resamples for each. In this evaluation, we use the 100 bootstrap samples for each independent dataset to construct 95% bootstrap percentile intervals (Efron & Tibshirani, 1993) and evaluate how often these intervals cover the true parameter value across 100 independent datasets.

Results

Table 1 shows the distribution of the percentage of bootstraps which select the correct number of classes for each criterion across the different simulation models. Without bootstrapping, the BIC selects the correct 2-class finite mixture model for all 100 independent datasets while the aBIC (94/100) yields slightly worse results. When examining each criterion across bootstrap samples, in only 44% of the independent datasets from a finite mixture model would the correct solution be chosen in more than 90 of the 100 bootstrap samples when using the BIC. At the same time, in 10% of the independent datasets the correct number of classes would have been chosen less than 80% of the time. Similar results, only worse, can be found for the aBIC.

The results for regression mixture models are worse still. In independent datasets the BIC selected the correct number of classes 100% of the time. However, when applied to bootstrap resamples from those independent datasets there was no case in which the BIC achieved 90% correct selection. In over 75% of the datasets the BIC selected the correct number of classes in less than 80% of the bootstraps. Results for the aBIC are mostly inferior still.

From these results it is apparent, that in contrast to individual datasets the non-parametric bootstrap increases the error made by the model selection criteria considerably. Although it is not reported in the table, in nearly all cases each selection criteria chooses too many latent classes, not too few. Would the bootstrap be used to verify the number of classes to use with these models, substantial error in the enumeration would be introduced and quite frequently the correct model would be rejected by the bootstrap despite the original solution being correct. Further investigations looking at the situation where a few outliers are present in the original data (not shown) yielded similar or even worse results. Using bootstrap to evaluate if a two-class finite mixture model is appropriate, when 5 outliers present shows that in 62% of the independent datasets the bootstrap chooses the 2-class model in less than 50% the resamples using the BIC to select the number of classes.

The second potential use of the non-parametric bootstrap has been to estimate the sampling distribution of parameters, if the correct 2-class model has been found in the original data. Table 2 shows the empirical coverage of bootstrap percentile confidence intervals as well as the empirical coverage of standard Wald-intervals of the relevant parameters for finite mixture and regression mixture models for the 2-class solutions. The coverage of the intervals is around the desired level of 95% when no outliers are present in the data as expected. This situation is, however, only of marginal interest as standard Wald type

intervals which are obtainable without the additional computational effort also yield adequate coverage (M. L. Van Horn et al., 2012). Once a small number of outliers are added to the data, the empirical coverage of the bootstrap confidence intervals deteriorates and in fact deteriorate at the same level as the coverage of the standard Wald intervals. For a finite mixture model, only 5 outliers (0.08% of the total dataset) are sufficient to yield coverage levels of only about 80% instead of the desired 95%. The results for the regression mixture models are a little bit better with 10 outliers resulting in coverage levels around 80–90% for some parameters. Thirty outliers, however, have a detrimental effect on the coverage. Despite the reasonable results for clean data, the bootstrap based intervals do not appear to add any robustness to the estimation of the confidence intervals of the model parameters.

A possible alternative

The above evaluation shows that the non-parametric bootstrap is not useful for validating results from mixture models and it is argued that the problem arises from replication of individual values that subsequently are regularly creating a new class. This section tests this hypothesis by evaluating a potential alternative, leave-k-out cross-validation (also called leave-k-out jackknife), which does not use sampling with replacement (Davison & Hinkley, 1997). This method takes subsamples of size $n-k$, where n is the total number of subjects in the study. As this approach does not allow for repeated sampling of the same value we anticipate that this approach will not encounter the same issue as the non-parametric bootstrap due to over-representation of single values. At the same time the fact that a reduced sample size is used might create problems with model fitting as it has been argued (M. Lee Van Horn et al., 2014) that mixture models are prone to extreme solutions when sample sizes become too small.

To decide on the number of observations to be left out, we consider two properties that are desirable: a) the size of the resampled datasets is large enough for stable model fitting, and b) in the presence of outliers, there is a small chance to resample these. We begin by considering that there are p outliers in an original sample of size n and denote the cross-validation sample of size by n_{cv} . Then the probability of having k outliers in a cross-validation sample can be found as,

$$P(k \text{ in } n_{cv}) = \frac{\binom{n-p}{n_{cv}-k} \binom{p}{k}}{\binom{n}{n_{cv}}}$$

Table 3 provides some of the above probabilities when 10 outliers are present in a dataset of 6010 for different sizes of the validation dataset. One can see that leaving 1000 observations out still leads to 98.4% of the datasets containing more than five outliers in the resample suggesting that leave-1000-out cross-validation is unlikely to allow robust estimation of the number of classes. Leaving 5000 observations out does yield a low probability of choosing too many outliers, yet in (M. Lee Van Horn et al., 2014)) it is argued that regression mixture models are a large sample technique. We therefore go with the middle ground of leaving half the data out in our evaluation below as this balances both considerations.

Like before we investigate the ability of standard model fit criteria (BIC, aBIC) to identify the correct number of classes in cross-validated samples. The same independent dataset as in the previous evaluation are used as the basis for the resampling method to ensure a fair comparison between the bootstrap and the leave-k-out cross-validation. We fit one, two and three-class models to 100 cross-validated datasets of size 3000 of each of the independent datasets and evaluate how frequently the correct 2-class solution is found. Table 4 shows the distribution of the percentage of cross-validation samples which select the correct number of classes for each criterion across the different simulation models.

The first notable point is that for both the finite mixture model and the regression mixture model without outliers all independent datasets choose the correct solution in more than 90 of the 100 cross-validated samples when using the BIC. These results are in stark contrast to the bootstrap (Table 1.) where 44% and 0% of the replications choose the correct solution using the BIC for finite mixture models and regression mixture models, respectively. From this we conclude that leave-k-out cross-validation does not suffer from the same problem of artificially adding additional classes as the bootstrap.

Following from this encouraging finding we also look at the situation when outliers are present. Again, the findings are encouraging in that for a large proportion of the independent datasets a large proportion of the cross-validated datasets choose the correct number of classes. For a regression mixture model with 30 outliers only 45 of the 100 independent datasets choose the correct 2-class model. For 73% of the independent samples, however, over 80% of the cross-validated datasets choose the 2-class model.

From the limited evaluation of the leave-k-out cross-validation it looks to be promising for validating class enumeration. When looking at its ability to appropriately estimate the sampling distribution of the parameter estimates (Table 5), however, it appears to be no different from the bootstrap. The performance of the cross-validation based percentile intervals is good without outliers. In the presence of outliers, however, the intervals show coverages below the desired level.

Discussion

Non-parametric bootstrap has been used for model validation and to obtain more robust estimation in finite mixture models in the past. The evaluations in this manuscript show that initial high hopes for the non-parametric bootstrap are not justified as it does not appear to help with stable selection of the number of components nor yield robust estimation in the light of very mild deviations from the model assumptions. Schlattmann (2005) argues that the non-parametric bootstrap can be used to obtain a consistent estimator for the number of components, we believe that an approach that only gets the right answer in 60–80% of the times for a sample size of 6000, is not useful in practice. As a consequence, we cannot recommend the non-parametric bootstrap for validating finite mixture model or regression mixture models. We attribute these results largely to the fact that influential individual observations have a high likelihood of being over-represented in the resamples. This feature, although an inherited property of the bootstrap, distorts the class enumeration of mixture models and impacts parameter estimation.

Using leave-k-out cross-validation for validating class enumeration results we find no artificial additional classes are introduced by this approach and that this approach shows promise even in the presence of outliers. The performance of the leave-k-out approach to estimate the sampling distribution of the parameter estimates, however, is no different from the bootstrap. One of the key challenges that will need to be answered comprehensively in further work before this approach can be recommended for validating enumeration results is the optimal size of the sub-sample. Taking a too small sample means that estimation will become unreliable while taking a sample that is too large will mean that outliers are likely to be included in most subsamples.

Acknowledgments

This research was supported by grant number R01HD054736, M. Lee Van Horn (PI), funded by the National Institute of Child Health and Human Development.

References

- Basford K, Greenway D, McLachlan G, & Peel D (1997). Standard errors of fitted component means of normal mixtures. *Computational Statistics*, 12(1), 1–18.
- Bauer DJ, & Curran PJ (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8, 338–363. [PubMed: 14596495]
- Bauer DJ, & Curran PJ (2004). The integration of continuous and discrete latent variable models: Potential problems and promising opportunities. *Psychological Methods*, 9, 3–29. [PubMed: 15053717]
- Davison AC, & Hinkley DV (1997). *Bootstrap Methods and Their Application*. New York: Cambridge University Press.
- de Jong MG, & Steenkamp JE (2010). Finite Mixture Multilevel Multidimensional Ordinal IRT Models for Large Scale Cross-Cultural Research. *Psychometrika*, 75(1), 3–32.
- DiCiccio T, & Efron B (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–228.
- Efron B, & Tibshirani R (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.
- George MRW, Yang N, Van Horn ML, Smith J, Jaki T, Feaster DJ, & Maysn K (2013). Using regression mixture models with non-normal data: Examining an ordered polytomous approach. *Journal of Statistical Computation and Simulation*, 83(4), 757–770. [PubMed: 23687397]
- Grün B, & Leisch F (2004). Bootstrapping finite mixture models. *Proceedings in Computational Statistics*, 1115–1122.
- Grün B, & Leisch F (2006). Finite mixture model diagnostics using the parametric bootstrap. Paper presented at the Junior scientist conference.
- Grün B, & Leisch F (2008). FlexMix version 2: finite mixtures with concomitant variables and varying and constant parameters. *Journal of Statistical Software*, 28(4), 1–35. [PubMed: 27774042]
- Grün B, & Leisch F (2010). Finite mixture model diagnostics using resampling methods. Vignette for R package flexmix.
- Lanza ST, Kugler KC, & Mathur C (2011). Differential Effects for Sexual Risk Behavior: An Application of Finite Mixture Regression. *Open Family Studies Journal*, 4, 81–88. [PubMed: 23125880]
- Leisch F (2004). FlexMix: A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software*, 11(8), 1–18.
- Lenk PJ, & DeSarbo WS (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1), 93–119.
- Liu M, & Lin T (2014). A skew-normal mixture regression model. *Educational and Psychological Measurement*, 74(1), 139–162.

- McLachlan G, & Peel D (2000). *Finite Mixture Models*. New York: John Wiley & Sons, Inc.
- Muthén LK, & Muthén BO (2010). *Mplus* (Version 6). Los Angeles: Muthén & Muthén.
- Newton MA, & Raftery AE (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B (Methodological)*, 3–48.
- Schlattmann P (2003). Estimating the number of components in a finite mixture model: the special case of homogeneity. *Computational Statistics & Data Analysis*, 41, 441–451.
- Schlattmann P (2005). On bootstrapping the number of components in finite mixtures of Poisson distributions. *Statistics and Computing*, 15, 179–188.
- Thompson B (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, 61(4), 361–377.
- Thompson B (1995). Exploring the replicability of a study's results: Bootstrap statistics for the multivariate case. *Educational and Psychological Measurement*, 55(1), 84–94.
- Van Horn ML, Fagan AA, Jaki T, Brown EC, Hawkins JD, Arthur MW, . . . Catalano RF (2008). The use of multilevel mixture models for the identification of differential intervention effects in a community randomized trial. *Multivariate Behavioral Research*, 43(2), 289–326. [PubMed: 26765664]
- Van Horn ML, Jaki T, Masyn K, Howe G, Feaster D, Lamont A, . . . Kim M (2014). Evaluating differential effects using regression interactions and regression mixture models. *Educational and Psychological Measurement*. doi:10.1177/0013164414554931
- Van Horn ML, Jaki T, Masyn K, Ramey SL, Antaramian S, & Lemanski A (2009). Assessing Differential Effects: Applying Regression Mixture Models to Identify Variations in the Influence of Family Resources on Academic Achievement. *Developmental Psychology*, 45, 1298–1313. [PubMed: 19702393]
- Van Horn ML, Smith J, Fagan AA, Jaki T, Feaster DJ, Masyn K, . . . Howe G (2012). Not quite normal: Consequences of violating the assumption of normality in regression mixture models. *Structural Equation Modeling*, 19(2), 227–249. doi:10.1080/10705511.2012.659622 [PubMed: 22754273]

Figure 1. Histogram of one dataset of size 100 and a single non-parametric bootstrap resample of size 100.

Figure 2.
Histogram of the finite mixture model 5 outliers (a) and scatter plot of the regression mixture model with 30 outliers (b)

Table 1.

Percentage of bootstrapped datasets selecting the correct 2-class solution for each criterion

	0-10	11-20	21-30	31-40	41-50	51-60	61-70	71-80	81-90	91-100
Finite Mixture										
No outliers										
BIC (100)	0	0	0	0	0	1	1	8	46	44
aBIC (94)	1	0	2	5	16	29	38	9	0	0
Regression Mixture										
No outliers										
BIC (100)	0	0	0	0	0	5	28	45	22	0
aBIC (96)	0	3	10	38	28	15	6	0	0	0

Note: Values in parenthesis are the number of times the respective criterion selects the correct two-class model in 100 independent datasets.

Table 2.

Empirical coverage of 95% bootstrap percentile intervals for different models with and without outliers. Coverages in parenthesis show the empirical coverages obtained by standard Wald intervals.

Finite Mixture	μ_1	σ_1^2		μ_2	σ_2^2		π
No outliers	0.96 (0.94)	0.94 (0.95)		0.94 (0.95)	0.96 (0.96)		0.96 (0.96)
Five outliers	0.78 (0.77)	0.73 (0.77)		0.79 (0.80)	0.82 (0.80)		0.77 (0.79)
Regression Mixture	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	π
No outliers	0.96 (0.96)	0.95 (0.96)	0.94 (0.95)	0.94 (0.94)	0.94 (0.96)	0.95 (0.97)	0.97 (0.98)
10 outliers	0.96 (0.96)	0.89 (0.94)	0.86 (0.93)	0.82 (0.93)	0.94 (0.95)	0.95 (0.95)	0.94 (0.94)
30 outliers	0.81 (0.83)	0.61 (0.80)	0.55 (0.53)	0.52 (0.54)	0.92 (0.91)	0.87 (0.86)	0.80 (0.85)

Table 3.

Probability of observing 5 or fewer (more than 5) outliers in a cross-validation dataset with various cross-validation sample sizes. The original dataset consists of 6010 subjects with 10 outliers.

no of outliers	Size of cross-validation dataset		
	1000	3000	5000
5 or less	0.998	0.625	0.016
More than 5	0.002	0.375	0.984

Table 4. Percentage of cross-validated datasets selecting the correct 2-class solution for each criterion

	0–10	11–20	21–30	31–40	41–50	51–60	61–70	71–80	81–90	91–100
Finite Mixture										
<i>No outliers</i>										
BIC (100)	0	0	0	0	0	0	0	0	0	100
aBIC (94)	0	0	0	0	0	0	1	1	10	88
<i>5 outliers</i>										
BIC (61)	0	0	0	0	0	2	2	15	31	50
aBIC (10)	0	1	6	18	24	23	16	9	3	0
Regression Mixture										
<i>No outliers</i>										
BIC (100)	0	0	0	0	0	0	0	0	0	100
aBIC (96)	0	0	0	0	0	0	1	6	50	43
<i>10 outliers</i>										
BIC (98)	0	0	0	0	0	0	0	0	0	100
aBIC (76)	0	0	0	0	2	4	15	18	40	21
<i>30 outliers</i>										
BIC (45)	0	0	0	0	2	5	7	13	25	48
aBIC (10)	6	14	17	14	12	16	12	7	2	0

Note: Values in parenthesis are the number of times the respective criterion selects the correct two-class model in independent datasets.

Table 5.

Empirical coverage of 95% cross-validation percentile intervals for different models with and without outliers. Coverages in parenthesis show the empirical coverages obtained by standard Wald intervals.

Finite Mixture	μ_1	σ_1^2		μ_2	σ_2^2		Π
No outliers	0.96 (0.94)	0.96 (0.95)		0.96 (0.95)	0.96 (0.96)		0.96 (0.96)
Five outliers	0.81 (0.77)	0.79 (0.77)		0.84 (0.80)	0.82 (0.80)		0.81 (0.79)
Regression Mixture	β_{10}	β_{11}	σ_1^2	β_{20}	β_{21}	σ_2^2	Π
No outliers	0.95 (0.96)	0.96 (0.96)	0.96 (0.95)	0.95 (0.94)	0.95 (0.96)	0.98 (0.97)	0.96 (0.98)
10 outliers	0.97 (0.96)	0.89 (0.94)	0.88 (0.93)	0.84 (0.93)	0.94 (0.95)	0.96 (0.95)	0.92 (0.94)
30 outliers	0.85 (0.83)	0.63 (0.80)	0.52 (0.53)	0.57 (0.54)	0.93 (0.91)	0.86 (0.86)	0.79 (0.85)