Psychometrically skewed distributions lead to finding pseudoclusters when using mixture models:

Commentary on "Heterogeneity in children at risk of math learning difficulties" (Munez et al., 2023)

Enrico Toffalini

Department of General Psychology, University of Padova, Italy

February 21, 2024

In a recent study, Muñez, Bull, Lee, and Ruiz (2023) challenge the dimensional framework on learning difficulties suggesting that children with math learning difficulties (MLD) might cluster into two (or even three) qualitatively different types, rather than representing cases of a single homogeneous population. They fit a series of confirmatory factor analysis (CFA), latent profile analysis (LPA) and factor mixture models. We praise the authors for identifying mixture models as an ideal method for investigating the existence of subpopulations within a larger set of cases, thus unveiling hidden heterogeneity in what is otherwise traditionally considered as an undifferentiated condition. In a similar fashion, in their recent comprehensive review on the "transdiagnostic revolution", Astle, Holmes, Kievit, and Gathercole (2022) suggested using clustering methods as a main way for discovering true underlying neurocognitive types beneath the surface of traditional diagnostic categories in neurodevelopmental disorders. While such a data-driven approach should not be used in isolation, Astle et al. (2022) present it as a major complement to the toolbox of the researcher who investigates neurodiversity, alongside dimensional methods.

Clustering techniques, including those based on mixture models, imply some assumptions. Like other statistical methods, when assumptions are not met results might still be largely valid, but validity is not guaranteed. Many mixture models used in psychology imply multivariate normal distributions, although the normality assumption is highly unlikely to be met in our discipline (Micceri, 1989). Moreover, even when assumptions are fully met results might be incorrect. For instance, a lack of statistical power may imply that some parameters are not adequately modeled. If the ground truth is unknown, however, such risks remain concealed. Luckily, data simulation allows conducting a priori sanity checks to quantify these risks beforehand. In the case of clustering techniques, one could simulate data from a given multivariate distribution, with pre-specified features such as correlation, skewness, and kurtosis coefficients (based on a priori knowledge or on own data). If data are drawn from a single population but the chosen clustering method prevailingly favors solutions that involve multiple clusters, this is clearly a problem, and it suggests that another method should be used.

To clarify the above, Toffalini, Girardi, Giofrè, and Altoè (2022) showed that Gaussian mixture models, under different sets of specific conditions, tend to incorrectly detect multiple clusters despite data being drawn from a single population. Surprisingly, this happened even when the distributional assumption was met (i.e., data were generated from perfectly Gaussian multivariate distributions). Such an inflation of detected clusters occurred when the sample size was insufficient for modeling the existing covariance across indicators within the detected clusters (e.g., the sample size was medium or even large, but the correlations across multiple indicators were weak). Strikingly, under specific combinations of sample size, correlation coefficients, and number of indicators, it was virtually guaranteed that Gaussian mixture models would detect multiple latent classes/clusters where none existed. Another problem is that for true clusters to be correctly detected, their separation across clustering indicators must be so large (i.e., at least Cohen's $d = 0.8$ on several independent dimensions; see Tein, Coxe, & Cham, 2013) that it is unlikely that they could have gone undetected until data-driven analysis was conducted. Such large effect sizes on many orthogonal variables may even appear implausible in psychological research per se (Toffalini et al., 2022). In Muñez et al. (2023), the separations between the two detected clusters are rather large, with the intercepts differing by more than 1.0 *SD* in all three indicators of math ability, and mean differences close to 1.0 *SD* also in several other domain-general and domain-specific cognitive variables. That one cluster outperforms the other(s) in all variables simultaneously might also indicate that the number of detected latent classes is inflated due to correlations (Toffalini et al., 2022).

As a retrospective evaluation of the inferential risks in the research scenario presented by Muñez et al. (2023), we run a series of Monte Carlo simulations. Data simulation was based on the coefficients that the authors report in their descriptive statistics table and was conducted with the "semTools" package of R, which allows to simulate multivariate (non-)normal distributions. We took sample size ($N = 428$), correlations, skewness, and kurtosis coefficients for the three measures of math ability used for clustering (i.e., math fluency, math problem solving, and numerical operations). All correlations are quite strong (r in [0.52, 0.70]). One skewness coefficient is large (0.98 in math fluency), and another is moderately large (-0.68 in numerical operations). All kurtosis coefficients are moderately large (absolute values ≥ 0.50). Crucially, large skewness here is likely to reflect the task characteristics, rather than the existence of clusters. Math fluency has a mean value (14.19) that is much closer to the lower (0) than to the upper (48) bound, and the mean is only 1.37 *SD*s away from the lower bound. Also, the data-generating process is a binomial one (i.e., sum of dichotomous [correct/incorrect] responses to items), which does not lead to normal distributions. Commendably, the authors provided the Mplus script that they used for fitting models, so we could reuse the exact same code in the first set of analyses. As Mplus is a licensed, non-open-source software, which makes it difficult for others to freely replicate the results, however, we subsequently offer other simplified examples entirely run with the R free software (R Core Team, 2010) to further clarify our points, and we provide the code.

We simulated 10 datasets and fitted factor mixture models using the Mplus code provided by

Muñez et al. (2023). Models featuring 1, 2, and 3 latent classes were fitted. Non-invariant model alternatives were preferred to ensure maximum flexibility, and because these were the models chosen by the authors in their final selected solution. In all 10 iterations, BIC consistently favored a three-class solution. The two-class model always outperformed the one-class model (median ΔBIC = -134.7), and the three-class model always outperformed the two-class model (median ΔBIC = -28.26). In all but one case, likelihood ratio test also suggested that one-class hypothesis (H_0) should be rejected in favor of a two-class solution (ps < 0.001), while in five of these cases the two-class hypothesis (as H_0) was rejected in favor of the three-class solution (ps < 0.05). As a further sanity check, all 10 datasets were re-simulated with perfectly normal distributions (i.e., with all skewness and kurtosis coefficients set to zero in the data-generating process). In this second case, the BIC consistently (and correctly) favored the one-class solution (median ΔBIC = +17.94 in favor of the one-class over the two-class solution, and median ΔBIC = +19.50 in favor of the two-class over the three-class solution). This double check confirmed that the inflation of the number of detected clusters was due to non-normality. Caution should be used, however, as several warnings emerged that the latent variable covariance matrix was not positive definite in one or more classes, especially with the three-class solutions.

Further Monte Carlo simulations of Gaussian mixture models were run in R. The code is provided on GitHub at `https://github.com/EnricoToffalini/commentary_mixture_skewness`. The famous "mclust" package (Scrucca, Fop, Murphy, & Raftery, 2016) was used. Tested solutions featured 1 to 3 components (latent classes/clusters). BIC was used as the criterion for determining the optimal solution. First, we run 1,000 iterations simulating datasets (each with $N = 428$) featuring skewness and kurtosis as reported above. Note that, in all cases, we drew data from one single population (with no clusters in it). The one-class solution was never selected, despite being the correct one. In 41.4% of iterations the two-class solution was favored, and in the remaining 58.6% of iterations the three-class solution was favored. Then, we run another 1,000 iterations with the same N and correlations across variables, but with all skewness and kurtosis coefficients set to zero. In this case, the one-class solution was (correctly) selected as optimal in 100% of iterations.

To summarize, we showed that there is likely no evidence of hidden heterogeneity in children with MLD. While doing so, we took the chance to point to a larger problem concerning clustering individuals on psychological data. The ground truth of the data-generating process is (always) unknown with respect to the real data, yet we demonstrated that Muñez et al. (2023) had a high chance of detecting multiple clusters in their data even if no true latent classes existed within

MLD. The main issue, in the present case, lies in data not meeting the normality assumption. As explained above, non-normality probably reflects the characteristics of the tasks used, and it is a widespread problem in psychometrics (e.g., Micceri, 1989). When assessing achievement or cognitive abilities in developmental samples, sum scores are often computed. Sums of dichotomous (correct/incorrect) responses, unfortunately, feature skewness (unless averaging at exactly 50%) and are certainly non-normally distributed, due to mean and *SD* being non-independent in binomial processes. Preliminary computation of individual ability parameters with Item Response Theory (IRT) models, or even via predicted scores with Confirmatory Factor Analysis (CFA) with ordinal items, might be a preferred solution here to normalize scores.

Moving forward, we suggest that clustering techniques should always be handled with caution in psychological research. This is particularly critical as inference emerging from clustering methods cannot easily be corroborated by other sources of evidence, and ground truth remains unknown. No descriptive statistics can back up inference unless visual inspection of scatter plots clearly suggests multimodal distributions. Unfortunately, this is unlikely to be the case in our discipline: it would require extremely large effect sizes (e.g., Cohen's ds $\gg 1.0$) for clusters to visually emerge from plotting alone. As a solution, we have recommended running Monte Carlo simulations as an a priori sanity check to see if the chosen clustering methods provide valid results when the ground truth is known, considering the characteristics of the data. Example of R code has been provided for it. In addition to quantifying a "*type I* error of clustering", that is the risk of detecting multiple classes when in fact only one exist, the readers may want to quantify statistical power, that is the probability of detecting the correct number of classes when true clusters/latent classes exist (Tein et al., 2013). As shown by both Tein et al. (2013) and Toffalini et al. (2022), however, sufficient power requires effect sizes of about Cohen's d = 0.80 or above in many independent dimensions simultaneously, which is bordering on credibility in psychology.

Code and data availability

Code and data used in this article are fully available on GitHub at: `https://github.com/EnricoToffalini/commentary_mixture_skewness`

References

Astle, D. E., Holmes, J., Kievit, R., & Gathercole, S. E. (2022). Annual research review: The transdiagnostic revolution in neurodevelopmental disorders. *Journal of Child Psychology and Psychiatry*, *63*(4), 397–417.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological bulletin*, *105*(1), 156.

Muñez, D., Bull, R., Lee, K., & Ruiz, C. (2023). Heterogeneity in children at risk of math learning difficulties. *Child Development*.

R Core Team. (2010). R: A language and environment for statistical computing. *(No Title)*.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using gaussian finite mixture models. *The R journal*, *8*(1), 289.

Tein, J.-Y., Coxe, S., & Cham, H. (2013). Statistical power to detect the correct number of classes in latent profile analysis. *Structural equation modeling: a multidisciplinary journal*, *20*(4), 640–657.

Toffalini, E., Girardi, P., Giofrè, D., & Altoè, G. (2022). Entia non sunt multiplicanda... shall i look for clusters in my cognitive data? *Plos one*, *17*(6), e0269584.