The paper I have chosen to read can be found at this link:

<https://pubmed.ncbi.nlm.nih.gov/33165028/>

The paper is titled the *Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls*.

In the field of pulmonary and critical care medicine, an often-used clustering method is Latent Class Analysis / Latent Profile Analysis. For our project with Dr. Eric Morrell, we were asked to perform such clustering on Dr. Morrell’s dataset.

The *Practitioner's Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls* provides an overview of Latent Class Analysis, as the title of the paper implies. The paper defines Latent Class Analysis (LCA) as a finite mixture model, where the subgroups / clusters are the “latent” groups that are found through patterns in features / variables or “indicators”. The article goes on to define two types of LCA. Latent Class Analysis (LCA), which is a finite mixture model which uses all categorical indicators and Latent Profile Analysis (LPA) which uses all continuous indicators. The article importantly notes that they use LCA as an overall blanket term for LCA and LPA. This is important as common literature in the pulmonary and critical care field also apply LCA as a blanket term. Some notable examples are Calfee et al.’s paper on acute respiratory distress syndrome (Calfee et al. 2021) and Bhatraju et al.’s paper on acute kidney injury (Bhatraju et al. 2019).

The paper provides a quick definition on how LCA works under-the-hood. LCA works by assuming the observed distribution in an indicator is the result of a finite “mixing” of underlying distributions. In other words, the overall distribution is a compilation of subgroup distributions. The paper then goes on to state that while Bayesian methods / statistics are provided for in LCA, solutions in LCA are usually acquired through maximum likelihood estimates.

With the general principle of how LCA, the paper then goes on to compare how LCA differs from cluster analysis such as K-Means and Hierarchal clustering. An important difference the paper points out is that cluster analysis

“separates study units into different clusters, whereas LCA estimates the probability that a given study unit belongs to each of the different latent classes”

(Sinha et al. 2021)

This is an important difference as it demonstrates that cluster analysis is based on arbitrary distance measures to define clusters while LCA uses distributions of the indicators to define classes.

A further interesting note is that the article cites an additional study where they evaluated the accuracy of LCA and K-Means. The additional study had two methods to find classes in a situation where the true class classification was already known but censored in the analysis. The paper notes that this simulated study found that K-Means had a misclassification rate of around 4-times greater compared to LCA. Another advantage they point out of LCA is that LCA is capable of handling mixed data types (categorical and continuous), whereas clustering methods like K-Means can only handle one or the other. While it may seem overwhelmingly advantageous to use LCA, the paper notes that the primary disadvantage of LCA is that it is extremely computationally demanding.

The paper “ends” with a section on “Key Steps” for performing LCA. The paper specifically provides a 5 stepwise approach:

1. Generate Hypothesis
2. Data set-up
3. Estimate Models
4. Evaluate Models
5. Interpret Optimal Models

The paper emphasizes the need for quality study design and indication selection, further perpetuating the adage that garbage-in leads to garbage-out. The paper emphasizes the need for a clear explanation on why an indicator should be included in the model.

In important aspect of these key steps is the examples and descriptions on how to select the optimal number of latent classes / optimal model. The paper notes that in the case of LCA with mixed categorical and continuous indicators that the “elbow method” using the BIC worked best in identifying optimal classes. Another method the paper points out is the Lo, Mendell, Rubin (VLMR) test and a bootstrapped version which is used to compare models of k classes to k-1 classes. However, the paper notes that they found that the bootstrapped version consistently selects k classes over k-1 classes which they claim limits the value of the bootstrapped version. For the sake of our analysis with Dr. Morrell, we will most likely perform an “elbow test” as most of our indicators are continuous with a mix of categorical which aligns with the case where the BIC elbow method is optimal.

In the closing of the paper, which also happens to be the interpretation section of optimal models, the paper raises important warnings / pitfalls to keep in mind when performing LCA. One of the pitfalls is described as the “Salsa effect”. This effect is when indicators of each class in a model are parallel to each other which simply suggest that the classes are on different scales of severity (mild, medium, and hot – in the terms of Salsa hence the name). If this occurs, the paper emphasizes the need to interpret the results cautiously.

# References:

Bhatraju, P. K., Zelnick, L. R., Herting, J., Katz, R., Mikacenic, C., Kosamo, S., Morrell, E. D., Robinson-Cohen, C., Calfee, C. S., Christie, J. D., Liu, K. D., Matthay, M. A., Hahn, W. O., Dmyterko, V., Slivinski, N. S. J., Russell, J. A., Walley, K. R., Christiani, D. C., Liles, W. C., … Wurfel, M. M. (2019). Identification of Acute Kidney Injury Subphenotypes with Differing Molecular Signatures and Responses to Vasopressin Therapy. *American Journal of Respiratory and Critical Care Medicine*, *199*(7), 863–872. <https://doi.org/10.1164/rccm.201807-1346OC>

Calfee, C. S., Delucchi, K., Parsons, P. E., Thompson, B. T., Ware, L. B., Matthay, M. A., & NHLBI ARDS Network. (2014). Subphenotypes in acute respiratory distress syndrome: Latent class analysis of data from two randomised controlled trials. *The Lancet. Respiratory Medicine*, *2*(8), 611–620. <https://doi.org/10.1016/S2213-2600(14)70097-9>

Sinha, P., Calfee, C. S., & Delucchi, K. L. (2021). Practitioner’s Guide to Latent Class Analysis: Methodological Considerations and Common Pitfalls. *Critical Care Medicine*, *49*(1), e63–e79. <https://doi.org/10.1097/CCM.0000000000004710>