# Practical Advice on LCA with John Dziak

December 1, 2016

Speaker 1: Methodology Minutes is brought to you by The Methodology Center at Penn State.

Hello, and welcome to Methodology Minutes. With me today is John Dziak, Research Associate at The Methodology Center. John is also the lead software developer, currently, on our LCA project, and he's an LCA researcher.

John, thanks for being here.

John Dziak: Thanks very much. Hello, everybody.

Aaron Wagner: John, just to kick this off, can you tell us what LCA is for?

John Dziak: I think that Bethany Bray and Stephanie Lanza already said this, but it's a latent variable technique that tries to summarize the responses of categorical items by dividing people into latent categories. Mathematically what's going on is that you're trying to find a simple model that summarizes the patterns of relationships between categorical items.

In practice, what it often means is that you're trying to find a story about how different groups of people respond differently, in systematic ways, to the items.

Aaron Wagner: Can you give us an example of why someone would want to use LCA?

John Dziak: I think people are looking for empirical subgroups for symptoms or behaviors. Sometimes this is done very self-consciously to try to come up with a typology for diagnosis, but I've seen articles where people have eating disorder symptoms or psychological symptoms, and say, "Well, how many kinds of this are there according to the data, not according to just my judgment or the way it seems to me?"

I think that can be very useful. One caveat is that you're certainly measuring phenotypes, not genotypes. There's no guarantee that you're discovering something new, in a hidden way, about nature or about how the classes came about. It might be that you're just finding the most convenient ways to divide people's probabilities of responses on these items into broad groupings of people, but that's okay, too.

Aaron Wagner: Thanks, John. Could you give us some insight into how LCA works?

John Dziak: You need a model to do anything. The LCA mathematical model is based on the idea of local independence. It's not assuming that the items are all independent from each other, that whether you drink is independent of whether you got drunk. That's ridiculous.

It's assuming that within any given class, your responses to items are ... Any relationship between them is already accounted for if you knew that the person was in that class. So if you knew that the person was a drinker, then the only way that they would say that they didn't drink was by random chance. Or, if you knew that the person had depression, which particular depression symptoms they mentioned, there was not further relationship with them. There weren't any subtypes of depression that you didn't find.

Probably that local independence never really holds, but that's like everything else in statistics, really, if you have a large enough sample and look hard enough. Think of a multiple regression setting. If your sample size is large enough, you can get anything to be significantly related to anything, even if the effect size is tiny.

If it's true ... This is paraphrased from a famous statistician, Box, who said that "all models are wrong, but some models are useful". Even if it's true that you might need 15 classes to totally account for every relationship seen in your data, maybe three classes will still tell you everything you really need to know and can interpret.

Aaron Wagner: Going back to that quote from Box, how do people know when they've got the "right model"?

John Dziak: Well, it's hard, because I'm not trying to advocate total relativism.

Aaron Wagner: Absolutely.

John Dziak: I'm not trying to say, "Well, whatever model makes you happy, that's the correct model." That isn't science.

Aaron Wagner: Right.

John Dziak: Even if I said that, it wouldn't matter, because, fortunately or unfortunately, we have peer reviewers and other people who are here to keep us accountable.

Aaron Wagner: Gatekeepers.

John Dziak: And they're going to say, "Well, why did you pick three classes? Why didn't you pick five? What criteria or tests did you use?" It's true that the criteria and tests are available, but they sometimes disagree, and the indications they give you might or might not be useful in your setting.

I was talking, just before, about some people call it being overpowered in statistics. If your sample size is very large, then any reasonable model, you will be able to find something wrong with it.

Aaron Wagner: You can identify a model if you've got a billion people in it.

John Dziak: Well, more like, even if the five class model already tells you everything that you can interpret ... For instance, suppose that in your topic, there aren't really five latent classes. There's a continuum of low to high, or maybe two dimensions. Depending on your sample size, the fits criteria might say, "Well, you should say there's low, medium, high," or if your sample size is bigger, they might say you should say low, medium-low, medium-high, high, and so on.

Aaron Wagner: And so on, yeah.

John Dziak: Some people would say, "Well, in that situation, latent class analysis isn't appropriate, because you should find some way of modeling the latent variable more realistically." I really am not convinced by that argument because I think that it's really hard to know what is the true shape of the distribution of a latent variable. It may be philosophically naive to say, "If I just make my model fancy enough, I will find a true and real number of classes in the Universe for my question."

Aaron Wagner: Right. The reality is messy and you may not be able to map your model onto it in a perfect way, no matter what you do.

John Dziak: Yeah. That reminds me of a joke I read in Reader's Digest, like 20 years ago, where the comedian was saying, "I have a map on a one to one scale. I live in square C2." Either we need to do more research on scaled fit indices and relative fit or maybe we just need to tell people, as Linda Collins and Stephanie Lanza do tell them in their book, I think, sometimes you have to use your judgment, in addition to the bootstrap test, and the AIC, and the BIC.

Aaron Wagner: Okay. All of that, I think, is a really excellent prelude to the big question of this podcast, which is, if I've done an LCA, how do I know whether or not it worked?

John Dziak: Well, nobody can tell you whether you asked the right question, so that's why ... You don't need to know what the right answer is, but you have to have some story before you start as to why you picked those items. It isn't completely garbage in, garbage out, but most of our users know that, I think.

If you're a very beginner, it's tempting to say, "Cool, I'll just shovel my whole data set into this algorithm and it will come out with insights for me." Sometimes, simple is good.

Aaron Wagner: How do we know if the LCA worked, though? What are we looking for?

John Dziak: Well, we have measures of relative fit if you could see do I need more classes? You could compare different numbers of classes informally, using the Akaike information criteria and the Bayesian information criterion. Those weren't developed for mixture modeling and they're not guaranteed to be always right, but they're a rule of thumb that people seem to respect and it's better than nothing.

You can also use the bootstrap likelihood ratio test to compare, do I need to add one more class in order to have adequate fit. Again, adequate is in the sense of statistical significance, not practical significance. It's okay to use your judgment if you don't think that an extra class is giving you anything you can actually interpret.

Aaron Wagner: Meaningful information.

John Dziak: Sometimes, actually, it's really important to put your foot down because we encourage people to use multiple starting values and look at what we call "identifiability". It's not quite the same things as identifiability in the statistical sense, which basically means you have positive degrees of freedom and a model that makes sense. What we identifiability is ...

We're not asking, "Have you found the right answer for the population?" We are asking, "Have you even found the right answer for your sample?"

Aaron Wagner: An explanation that you had given me for this, at another time, that really stuck in my head, I'm going to see if I can reflect it back to you. Let me know if I'm off base here. You encouraged me to imagine the class structure as a three dimensional landscape, like a hilly area.

The algorithm, picking a random start, it puts somebody somewhere on the map. So you're somewhere in a valley, or somewhere, and then your little person, imaginarily, on the map, goes straight to the top of the nearest hill. But it's foggy up there and they can't tell whether this is the highest hill or just the hill they happen to start up.

They know they're on the top of something, but they can't tell whether it's a local maximum, just a short hill that they happen to be near, or the global maximum, the highest point in the entire mountain range.

John Dziak: That's exactly how it works.

Aaron Wagner: So, by starting, if you drop people a hundred times on the map and they keep getting to the same high point, odds are that that's-

John Dziak: Or even if half of them get to the same high point and there's a hundred of them, odds are that it's the highest. You can never prove that there might not be a flagpole somewhere that's almost impossible to find, but higher than everything else, but now you have to go back to common sense. If your situation is that bizarre and contorted, well, maybe there's a class for unicorns, but maybe it's okay if you don't find it.

Aaron Wagner: Yes, right. Exactly, exactly.

John Dziak: Yes, that's a good way to put it. We certainly encourage everybody to use end starts unless you have starting values that you've chosen and you really believe in. Personally, I think you should use end starts anyway, because just because you believe something doesn't mean it's true.

Aaron Wagner: Yeah. You think you're putting the person at the base of the highest mountain on the map, but-

John Dziak: If you really knew that-

Aaron Wagner: Why are you doing Latent Class Analysis?

John Dziak: Why do you need to explore at all? However, that's not what we said 10 years ago and that's not what some smart people would still say today. They would say, "This is supposed to be about theory, so you should use your theory at every stage."

You should definitely use either end starts or a theoretically driven starting value data set. You should never just say, "Start. One, two, three, four, five, and go." Which one of those two good options is the best is, like a lot of other things, a judgment call.

We like to see what we call "percent identified". Again, there's no cut off.

Aaron Wagner: Yeah, no heuristic you can really follow.

John Dziak: Yeah, we don't really say it has to be 20% or it has to be 50%, but if it's like 5%, it's somewhat worrisome. If AIC or BIC is telling you to have an eight class model, but your answers for the eight class model are so unstable that they're practically random, you pretty much have a responsibility to use a smaller number of classes, because you don't want to be publishing random results.

Aaron Wagner: It sounds like you're saying the way to know if it worked is to look at your fit statistics, look at your relative identifiability, and to understand your theory and apply it to the information that you're getting out of the model.

John Dziak: Yeah, I'm certainly not saying that you know you've got the right answer if it was the answer you expected. That's not science.

Aaron Wagner: Right. No.

John Dziak: But if it's something you can't explain at all, then it probably won't do you or your readers any good, and it will quite possibly lead people astray because they'll start coming up with fanciful stories for why is it that way?

Aaron Wagner: How do I pick my items for my model?

John Dziak: At least our implementation of LCA seems to work best when you have a fairly moderate amount of items, like maybe 10 or less, each with a small number of categories, probably two would be the best number, and you know what each item means.

What we call PROC LCA is not really intended for mining large data sets with hundreds of variables. First of all, our software won't converge with that, and second, it's not what our software was designed for.

Items kind of get an equal vote once you let them in the model. As it is, with factor analysis, also, if you have 10 items on alcoholism and two items on smoking, the class structure that's found will have more to do with alcohol than about smoking. That's not a discovery. It's a consequence of the way you constructed your model.

Co variates sometimes change the meaning of your classes. Items always do. When you're choosing measurement items, I guess this is just common sense, but they should all have to do with the story you're trying to tell or the question you're trying to investigate. That's more important than just putting in a lot of data and hoping something good comes out.

Aaron Wagner: Let's talk now about selecting the model.

John Dziak: Just like anything else in statistics, you want the model to be rich enough to tell you most of what's going on in your data set, but not so rich that it includes a lot of noise. How do you tell whether you have enough classes but not too much?

One way is to use AIC and BIC. Another way is to use the bootstrap likelihood ratio test. Bethany Bray and Stephanie Lanza discussed both of these in podcast 15. As we mentioned before, another important thing to consider is the identifiability, because your fit statistics ... It's pretty common. It's pretty common that if you have a large sample your fit statistics would say, "Well, you know, I could make the fit even better if you let me have 10 classes instead of nine or 11 instead of 10," but then when you check the percent identifiability, it's very poor.

We don't have a rule for you to follow, like in high school math, that will automatically get you the right answer, but you have to balance these two things in your mind somehow and probably the tie breaker will be the interpretability of the solution.

You'll print out the estimates from each of the model sizes you're considering and try to find out whether they tell you a story that makes sense. There might not be a right answer. There might be several useful answers, depending on what question you're asking.

Aaron Wagner: There is a trade off.

John Dziak: Yeah, yeah, and in practically any kind of statistical model, there's something called a bias variance trade off. If you pick a simple model, then it will be somewhat biased. Not in the sense of it has an ax to grind, but in the sense of there are some features it's just going to miss because it wasn't looking for them.

However, if you pick a complicated model, even though you're increasing your potential to tell the right story, you're also much increasing your risk of telling a totally wrong story because of sampling error, or random variability, or you just misunderstood the answer. Somewhere in between is a happy medium.

Finding it is not really ... There are tools to try to help you find it, like AIC and BIC, but they all have their own assumptions and none of them are infallible, so finding it is kind of the main thing that makes these analysis an art and not just a science.

Aaron Wagner: Can you tell us what percent identifiability we should look for when selecting a model?

John Dziak: First of all, if you use a really low number of random starts, then I can't give you any heuristic, because if you use, let's say, three, they might all agree by chance. Let's say you use 100. I think if you're just exploring, you're checking out different model sizes, you're probably going to use like 20, but let's say you use 100.

I think that if you're using 100, to me, I would feel okay if at least 25% of them got to the best. I wouldn't necessarily say it was wrong if less than 25% got you to the best obtained answered, but if 25 times, they were able to find it and all the other times were something poorer, it seems like there's probably not something that much better to be easily found. That would seem reasonable to me.

Aaron Wagner: When should a user bump up the number of end starts they're using?

John Dziak: I would like them to always use at least 20. Maybe you would increase it if the percent agreement was kind of low and so you wanted to see if, "Maybe if I dropped more people onto that landscape, maybe some of them would find a taller mountain." If it's just not working and you're trying to increase end starts so that you can get some kind of an answer eventually, that's probably not as good.

Aaron Wagner: Right. So questionable percentage of seed agreement would be a reason that you would encourage people to bump it up?

John Dziak: Yeah. It's that you don't feel it is awful but you don't feel it is great.

Aaron Wagner: Let's say I'm using PROC LCA and my model won't converge. What causes that and what should I do about it?

John Dziak: Sometimes that can be fixed by adding a small prior, by adding a prior on Beta or Rho. Sometimes you don't converge because one of the co variates has a odds ratio that's extremely high and it can't find a finite estimate for it.

I guess a classic example would be one of the classes is pregnant and one of the co variants is biological sex, so the males just can't get pregnant, with any probability, so their odds ratio would be infinity. If any of the estimates come out as infinity, the PROC will tell you it didn't converge, because converging means comparing to something finite.

In that case, either you would look at your items and use common sense or I guess you could use a prior. The prior is usually able to keep all the estimates finite. Keep it small, like prior size of one or .1.

Another reason that your model might not converge is that you might just be asking for more than the data can give you. Let's say you have eight three-level items. That means that the contingency table that you're trying to describe is a table with the three to the eighth different cells and you're trying to find a model to predict the proportions in each of those cells.

Some of those cells are going to have really, really small proportions. It's going to be really hard to even approximate them all. The word converged means that the algorithm was able to find one set of estimates that gives you a better likelihood than any of the others that are in the neighborhood.

If the model is so rich and the data is so poor that there are many ways to represent the data about equivalently, then either it won't converge or it will converge just to a local maximum because there is no global maximum, or the global maximum is found in many places.

If your model ... My favorite thing to tell people, I don't think it's their favorite thing to hear, is that if your model doesn't converge, use fewer co variates or fewer classes. That sounds like throwing out information and maybe it sounds unscientific, and maybe, "Well, the AIC told me to use eight classes," but again, having a stable and interpretable answer is just as important as having a extremely thorough answer.

Another way to help models converge is to use the restrict option, but I don't encourage people to use it unless they have a really good a priori reason and know what they're doing. A classic example of the restrict option would be the binge drinking class has zero probability of not using alcohol at all because if you're a total abstainer, then there's no way of binge drinking.

If you're just putting in restrictions to see which ones will get you to converge, that's not what restrict is designed for, because then you're putting in a posteriori restrictions as though they were part of the definition of your model. You're putting them in as a priori, even though they're really exploratory.

Partly because restrict is hard use and easy to make a mistake with, but partly because most of the time, I think we don't have a definite enough idea of what restrictions we want in order to make it scientifically reasonable to use restrict.

I usually like to try to get people to simplify their model first. It sounds like a mean thing to do because, "Well, I could learn so much more if I had a richer model," but if make your model too rich for you to understand or for your data to identify, then you can start learning stuff that ain't so.

Aaron Wagner: Right. John, thank you very much for being with us today.

John Dziak: Oh, you're welcome. Thanks very much for talking with me.

Aaron Wagner: You have been listening to Methodology Minutes. Please visit methodology.psu.edu for podcasts, software, reading lists, and more.