

Bijlage A

Toelichting op de latente klassenanalyse

Met latente klassenanalyse (LCA) kunnen in een populatie groepen van gelijksoortige eenheden worden gevonden op basis van hun kenmerken op een aantal relevante variabelen (van categorisch, ordinaal en/of interval niveau). “Latente klassen” kunnen worden opgevat als “verborgen groepen”. Per analyse dient het aantal latente klassen aangegeven te worden. Het is gebruikelijk om het aantal latente klassen te laten variëren van twee tot bijvoorbeeld tien (of meer indien nodig). Op basis van een aantal statistische maten en de interpreteerbaarheid van de klassen kan een beslissing genomen worden over welk aantal latente klassen de voorkeur verdient. De latente klassen laten zich vervolgens beschrijven op basis van hun profiel op de variabelen.

Omdat de onderhavige data longitudinaal zijn is het latent class growth analysis (LCGA) (in Mplus) geschikt om patronen in de loopbanen van managers te vinden. De LCGA is uitgevoerd op de acht waves van de variabele die aangeeft hoeveel ondergeschikten een manager heeft (een selectie van minstens vijfmaal antwoord gegeven resulteert in $n = 2085$ respondenten, inclusief degenen die nooit manager zijn geweest). In de eenvoudigste LCGA wordt alleen een lineair effect over de waves gemodelleerd, maar hieronder wordt gebruik gemaakt van een lineaire plus een kwadratische component, waarmee kromlijnige (curvilineaire) effecten kunnen worden waargenomen.

Hoe bepaal je het optimale aantal latente klassen?

Het Bayesiaanse Informatie Criterium (BIC¹) wordt vaak gebruikt om het optimale aantal latente klassen te bepalen. Een “informatiecriterium” is een statistische maat gebaseerd op een aantal statistische gegevens van de uitkomsten van een analyse. Het BIC is een maat die strenger is dan het AIC (Aikaiken Informatie Criterium) om *overfitting* te voorkomen – het BIC streeft naar zuinige modellen. Een alternatief is het *sample size adjusted* BIC² (SABIC), wat qua strengheid tussen het BIC en het AIC in zit. Voor zowel BIC als SABIC geldt: hoe lager de waarde, hoe beter de fit (hoe beter het model past bij de data).

Er zijn ook speciale toetsen ontwikkeld om het optimale aantal klassen te bepalen. De Lo–Mendell–Rubin toets (LMR) toetst of een oplossing met een bepaald aantal latente klassen significant beter is dan de oplossing met één klasse minder. Als deze toets **niet** significant is (dus $p > 0,05$), dan is de oplossing met één klasse minder een optimale oplossing. Dezelfde procedure geldt voor de Bootstrap Likelihood Ratio Test (BLRT).

Nylund, Asparouhov en Muthén (2007) concluderen in hun Monte Carlo-onderzoek met diverse modellen en steekproefgroottes, dat er sterk bewijs lijkt te zijn dat de BIC de beste statistische maat is voor het bepalen van het aantal latente klassen. Van de toetsen is de Bootstrap Likelihood Ratio Test (BLRT) beter dan de LMR. Over het geheel genomen presteert de BLRT beter dan de BIC.

¹ De BIC hecht veel waarde aan spaarzaamheid van het model. De BIC wordt strenger naarmate de steekproefomvang toeneemt: $-2\log\text{-likelihood} + p * \ln(n)$ waarbij $\ln(n)$ het natuurlijke logaritme van de steekproefomvang is en p het aantal vrije modelparameters is.

² Sample-size adjusted BIC: Net als bij het BIC, wordt het SABIC strenger bij het toevoegen van parameters op basis van de steekproefomvang, maar minder streng dan het BIC. Vervang in de formule van BIC de n door: $(n + 2) / 24$

Verder kan nog de entropie van een latenteklassenanalyse worden bepaald. Dit wordt niet als een fitmaat beschouwd, maar geeft wel aan hoe adequaat latente klassen van elkaar zijn te onderscheiden. Hoe dichter de entropie bij één ligt, des te beter is het onderscheidingsvermogen. Hoe dichter bij nul, des te meer overlappend en onduidelijk de latente klassen zijn.

In tabel A.1 staat relevante informatie over LCGA's van een tot en met acht latente klassen. De Lo–Mendell–Rubin toets (LMR) is niet-significant bij zes latente klassen, hetgeen betekent dat de LCGA met vijf latente klassen volgens dit criterium het uitverkoren model is. De BLRT blijft steeds significant, dit is een bekend probleem waar al veel onderzoekers op het internet over klagen. De praktische bruikbaarheid is hierdoor niet erg groot. De BIC en SABIC blijven dalen naarmate het aantal latente klassen toeneemt en geven hier dus geen uitsluitsel. Wel is te zien dat de BIC's na de oplossing met vijf latente klassen nog maar weinig afnemen.

Tabel A.1 Fitmaten van LCGA's met lineaire en kwadratische groeicurve (n = 2085) naar het onderscheiden aantal latente klassen

Aantal latente klassen	Log-likelihood	Df	BIC	SABIC	Entropie	LMR	$P \leq$	BLRT	$P \leq$
1	-14521	8	29102	29077	-	-	-		
2	-12014	12	24120	24082	0,91	4854,05	0,001	5012,83	0,001
3	-11382	16	22886	22835	0,90	1224,93	0,001	1265,00	0,001
4	-11242	20	22637	22574	0,90	270,22	0,001	279,06	0,001
5	-11132	24	22448	22371	0,91	286,41	0,012	295,78	0,001
6	-11074	28	22361	22272	0,88	136,96	0,388	141,44	0,001
7	-11053	32	22350	22248	0,89	142,22	0,300	145,32	0,001
8	-11016	36	22306	22192	0,88	23,53	0,509	24,04	0,001

BIC: Bayesiaans Informatie Criterium.

SABIC: sample size adjusted BIC.

LMR: de Lo-Mendell-Rubin toets met in de kolom ernaast de bijbehorende p -waarde.

BLRT: de Bootstrap Likelihood Ratio Test met in de kolom ernaast de bijbehorende p -waarde.

Latente Klasse Groei Analyse op respondenten die 35 jaar of jonger waren bij de eerste wave

Tabel A.2 Fitmaten van LCGA's met lineaire en kwadratische groeicurve (n = 550) naar het onderscheiden aantal latente klassen

Aantal latente klassen	Log-likelihood	Df	BIC	SABIC	Entropie	LMR	$P \leq$
2	-3139	12	6353	6315	0,91	1161,96	0,001
3	-2963	16	6028	5977	0,89	337,40	0,001
4	-2925	20	5976	5912	0,90	74,16	0,519
5	-2898	24	5947	5871	0,89	51,63	0,260
6	-2889	28	5954	5865	0,79	37,76	1,000
7	-2869	32	5941	5839	0,87	48,89	0,052

In tabel A.2 staan de fitmaten van de LCGA's op de groep jonge respondenten. Volgens de Lo, Mendell, Rubin test is de latenteklassenoplossing met 4 klassen niet significant beter dan die met 3 klassen, en wordt dus de 3 klassenoplossing uitverkoren. Het BIC-criterium is het laagst bij 5 latente klassen, maar dit is praktisch niet haalbaar omdat de twee kleinste klassen dan slechts zeer weinig respondenten bevatten.

Latente Klasse Groei Analyse op respondenten ouder dan 35 jaar bij de eerste wave

Tabel A.3 Fitmaten van LCGA's met lineaire en kwadratische groeicurve (n = 1264) naar het onderscheiden aantal latente klassen

Aantal latente klassen	Log-likelihood	Df	BIC	SABIC	Entropie	LMR	$P \leq$
1	-9120	8	18297	18272			
2	-7395	12	14876	14838	0,925	3332,74	0,001
3	-6963	16	14040	13989	0,915	836,09	0,001
4	-6851	20	13844	13781	0,912	216,59	0,016
5	-6754	24	13680	13604	0,923	185,71	0,003
6	-6739	28	13677	13588	0,93	189,00	0,004
7	-6664	32	13557	13456	0,872	62,62	0,187

In tabel A.3 staan de fitmaten van de LCGA's op de groep oudere respondenten. Volgens de Lo, Mendell, Rubin test is de latenteklassenoplossing met 7 klassen niet significant beter dan die met 6 klassen. Ook het BIC-criterium is het laagst bij 7 klassen (al zou dit nog lager kunnen zijn bij 8 klassen). Maar 7 klassen resulteert in een kleinste latente klasse met slechts drie respondenten wat duidelijk te weinig is. De oplossing met vijf latente klassen geeft nog redelijke aantallen en wordt daarom gekozen.