hyväksymispäivä arvosana

arvostelija

Kausaaliset kaaviot empiirisessä tutkimuksessa

Jarkko Salminen

Helsinki 18. maaliskuuta 2003

Kausaalimallien tutkimusseminaari

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Sisältö

1 Johdanto

Empiiristen tutkimusten tavoitteena on selvittää syy-seuraus -suhteita. Johtopäätökset kausaalisista suhteista pyritään tekemään tilastollisen aineiston perusteella. Kausaalisten mallien teorian mukaan pelkän tilastollisen datan tarkastelu ei kuitenkaan riitä, vaan oikeiden johtopäätösten tekemiseksi on tunnettava kausaalinen tarina, joka kertoo, miten aineisto on saatu.

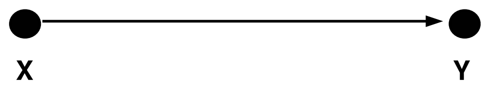
Tilastotiede tarjoaa syy-seuraus -suhteiden tutkimiseen satunnaisesti kontrolloitua koetta (randomized controlled experiment), joka on sen näkemyksen mukaan ainoa luotettava menetelmä päätellä kausaalisia seurauksia. Satunnaisesti kontrolloidussa kokeessa perusjoukosta valitaan satunnaisotannalla otos, jonka on tarkoitus edustaa perusjoukkoa. Näin otoksen avulla pyritään tekemään induktiivisia johtopäätöksiä perusjoukosta. Saadusta otoksesta valitaan edelleen satunnaisesti koe- ja kontrolliryhmä. Koeryhmä altistetaan koekäsittelylle, ja kontrolliryhmälle käsittelyä ei tehdä. Sitten koeryhmän tuloksia verrataan kontrolliryhmään, jotta saadaan selville käsittelyn vaikutus.

Määritellään seuraavaksi tarkemmin mitä tarkoitetaan kausaalisella seurauksella, jota seuraavissa luvuissa on tarkoitus estimoida. Määritelmän 1 mukaan muuttujan X kausaalinen seuraus muuttujalle Y on todennäköisyysjakauma $P(y|do(x))$, joka antaa todennäköisyyden Y:n arvolle y, kun interventiolla X:lle asetetaan arvo x. Tarkastelun tavoitteena on siis laskea jakauma $P(y|do(x))$ tilastollisen aineiston ja kausaalisten kaavioiden avulla.

Määritelmä 1 (Kausaalinen seuraus) *Olkoon X ja Y kaksi erillistä kausaalisen mallin muuttujajoukkoa. X:n kausaalinen seuraus Y:lle on todennäköisyysjakauma $P(y|do(x))$.*

Merkitään nyt muuttujalla X sitä, tehdäänkö kontrolloidussa kokeessa käsittely yksittäiselle koeyksikölle, ja muuttujalla Y seurausta. Tällöin kokeen toden-

Kuva 1:



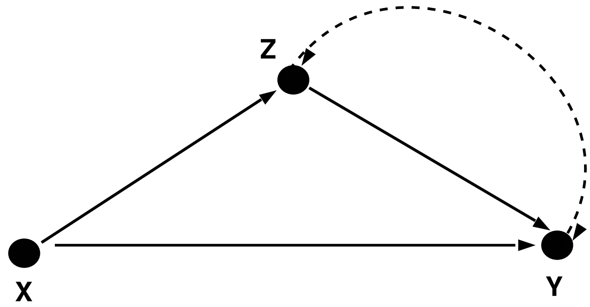
X ja Y ovat sekoittumattomia.

näköisyysjakauma $P(y|x)$ voidaan tulkita kausaaliseksi seuraukseksi: todennäköisyys seuraukselle y, kun altistetaan mikä tahansa yksikkö käsittelylle x. Toisin sanoen havaitaan, että satunnaisesti kontrolloidussa kokeessa $P(y|x)$ on yhtä kuin $P(y|do(x))$. Tämä johtuu siitä, että kokeessa valitaan satunnaisesti, tehdäänkö koeyksikölle altistus. Silloin kokeen tilastollinen aineisto perustuu kausaaliseen malliin, jossa mikään muuttuja ei voi vaikuttaa X:ään. Mallin kaaviossa 1 X:ään ei osoita yksikään nuoli eikä interventio poista kaaviosta yhtään X:ään osoittavaa nuolta (interventio X:lle tarkoittaa X:ään osoittavien nuolien poistamista kaaviosta). Näin ollen interventio ei muuta mallia, ja kausaalinen seuraus $P(y|do(x))$ saadaan suoraan jakaumasta $P(y|x)$.

Samaan huomioon päädytään myös sekoittumattomuuden käsitteen kautta. Koska X:llä ja Y:llä ei ole yhteisiä esi-isiä kontrolloidun kokeen kaaviossa 1, X ja Y ovat sekoittumattomia. Sekoittumattomuuden määritelmän mukaan silloin $P(y|do(x))$ on sama kuin $P(y|x)$.

Satunnaisesti kontrolloitu koe on siis luotettava tapa arvioida kausaalista seurausta, koska se estää sekoittumisen. Tällaista koetta ei kuitenkaan aina voida tehdä: syynä voi olla tutkimuksen korkeat kustannukset tai eettiset kysymykset. Silloin voidaan vain kerätä havaintoja ja yrittää estimoida kausaalista seurausta saadun aineiston avulla. Ongelmana tällaisessa havaintotutkimuksessa (observational study) on kuitenkin sekoittumisen riski, minkä vuoksi tilastotieteessä varoitetaan tekemästä johtopäätöksiä syy-seuraus -suhteista havaintotutkimusten perusteella ($P(y|do(x))$ voi olla erisuuri kuin havaittu $P(y|x)$).

Kuva 2:



X ja Y ovat sekoittumattomia.

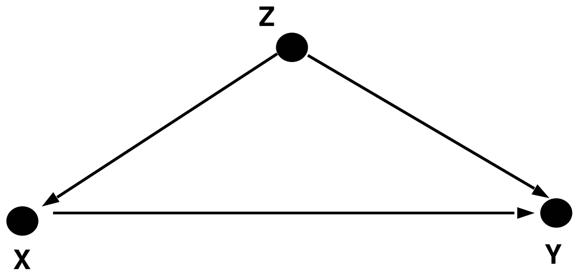
Kausaalisten mallien avulla sekoittumista voidaan kuitenkin hallita havaintotutkimuksissa. Kuvaamalla aineiston alkuperä kausaalisella kaaviolla seuraus $P(y|do(x))$ voidaan johtaa havaitusta jakaumasta.

2 Sekoittumisen hallinta havaintotutkimuksissa

Kausaalinen seuraus voidaan havaintotutkimuksessa laskea suoraan sekoittumattomuuden määritelmästä, jos ollaan täysin varmoja, että X:llä ja Y:llä ei ole yhteisiä esi-isiä (kaaviot 1 ja 2). Sovellusaluekohtaisen tietämyksen nojalla on mahdollista tietää, ettei ole olemassa mitään yhteistä syytä.

Jos kuitenkin muuttujat X ja Y ovat sekoittuneet, kausaalinen seuraus voidaan silti laskea, jos tietyistä mallin muista muuttujista on tehty havaintoja. Taka- ja etuovikriteerien avulla kaavioista pystytään löytämään näitä muuttujia. Esimerkiksi kuvassa 3 takaovikriteeri takaa, että seuraus on laskettavissa, koska Z:sta on havaintoja. Kaavioiden merkinnöissä tummat ympyrät tarkoittavat muuttujaa, josta on tilastollista aineistoa (havaittu), ja vaaleista ei ole havaintoja (havaitsematon).

Kuva 3:



X:n kausaalinen seuraus Y:lle on identifioituva.

Seurauksen laskettavuudesta käytetään nimitystä identifioituvuus. Jos seuraus on identifioituva, se voidaan laskea havaittujen muuttujien jakaumasta (määritelmä 2).

Määritelmä 2 (Identifioituvuus) *Muuttujan X kausaalinen seuraus muuttujalle Y on identifioituva, jos todennäköisyysjakauma $P(y|do(x))$ voidaan laskea yksikäsitteisesti mistä tahansa havaittujen muuttujien positiivisesta jakaumasta.*

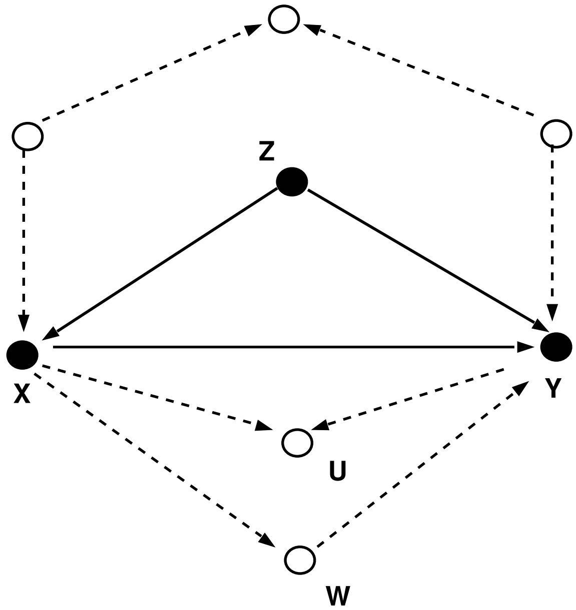
Tavoitteena on siis tutkia, onko kausaalinen seuraus identifioituva tietyssä mallissa, kun vain osasta muuttujia on dataa. Jos se ei ole identifioituva, ovikriteerien avulla kaaviosta nähdään, mitä lisähavaintoja tutkimussuunnitelmaan on lisättävä, että seuraus voidaan laskea.

2.1 Takaovikriteeri

Takaovikriteeri on graafinen testi, joka kertoo, onko havaittu riittävästi muuttujia kausaalisen seurauksen identifioimiseksi (määritelmä 3).

Määritelmä 3 (Takaovikriteeri) *Joukko Z muuttujia toteuttaa takaovikriteerin suhteessa järjestettyyn pariin muuttujia (X_i, Y_j) suunnatussa syklittömässä kaaviossa G, jos*

Kuva 4:



Muuttujan Z havaitseminen toteuttaa takaovikriteerin.

(i) mikään Z:n solmuista ei ole X_i:n jälkeläinen ja

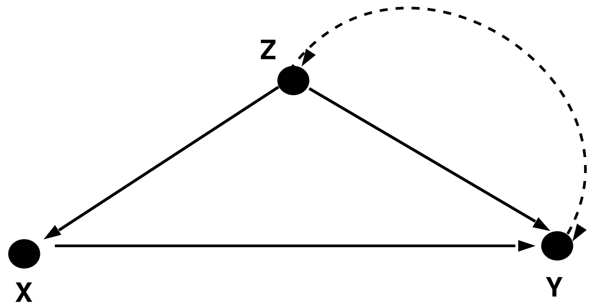
(ii) Z d-erottaa kaikki X_i:n ja Y_j:n väliset polut, jotka sisältävät nuolen muuttujaan X_i.

Jos vastaavasti X ja Y ovat kaksi erillistä G:n solmujoukkoa, niin silloin Z toteuttaa takaovikriteerin suhteessa järjestettyyn pariin (X, Y), jos Z täyttää kriteerin suhteessa kaikkiin pareihin (X_i, Y_j), kun $X_i \in X$ ja $Y_j \in Y$.

Kaavio 4 selittää kriteerin määräämät lisähavainnot. Se esittää kaikkia X:n ja Y:n välisiä mahdollisia polkuja. Malliin on merkitty tummalla muuttuja Z, joka takaoviehtojen perusteella on havaittava. Polkuja, jotka alkavat nuolella X:ään, kutsutaan takaovipoluiksi. Ehdon (ii) mukaan on riittävää tarkastella vain näitä takaovipolkuja, mistä kriteeri saa nimensä.

D-erotuksen mukaan takaovipoluilla on oltava yksi havaittu muuttuja tai törmäävät nuolet. Tätä esittää kaaviossa kaksi ylimmäistä polkua: yhteisestä syystä on havaintoja (muuttuja Z), tai sitä ei ole olemassa (ylin polku). Mallista on jätet-

Kuva 5:



Kausaalinen kaavio, jossa takaovikriteeri pätee

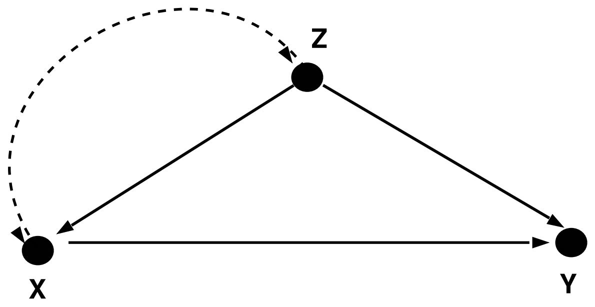
ty pois sellaiset takaovipolut, jotka päättyvät Y:stä lähtevään nuoleen, koska ne sisältävät joko törmäyksen tai aiheuttavat kaavioon syklin.

Takaovettomilta poluilta ei tarvita havaintoja. Tätä kuvaavat mallissa kaksi alinta polkua. Näillä reiteillä on välttämättä törmäys (muuttujassa U), jos Y:stä lähtee nuoli, tai yhteistä syytä ei ole, koska X:ään ei osoita nuolta (alimmainen polku). Näin ollen kaksi alimmaista polkua ja niiden muuttujat ovat merkityksettömiä, ja ne voitaisiin poistaa mallista.

Sen sijaan on tärkeää, että kaaviossa on mukana kaikki mahdolliset nuolet mallin muuttujien välillä eikä myöskään yhtään yhteistä syytä saa jättää pois. Muuten seurauksen estimaatista saattaa tulla harhainen. Esimerkiksi kaaviot 5 ja 6 toteuttavat takaoviehdot, mutta yhdenkin nuolen tai kaaren lisääminen rikkoisi kriteerin. Mallien katkoviivaiset kaaret, joissa on nuoli molemmissa päissä, tarkoittavat, että muuttujien välillä on polkuja, jotka sisältävät vain havaitsemattomia muuttujia eivätkä lainkaan törmääviä nuolia. Silloin muuttujilla on yhteinen esiisä, ja ne ovat sekoittuneita.

Jos takaovikriteeri pätee, kausaalinen seuraus on identifioituva, ja se saadaan kaavasta 1.

Kuva 6:



Kausaalinen kaavio, jossa takaovikriteeri pätee

Teoreema 1 *Jos joukko Z muuttujia toteuttaa takaovikriterin suhteessa järjestettyyn pariin (X, Y), X:n kausaalinen seuraus Y:lle on identifioituva ja saadaan kaavasta 1.*

$$P(y|do(x)) = \sum_{z} P(y|x, z)P(z) \tag{1}$$

2.2 Etuovikriteeri

Jos kausaalista seurausta ei pystytä toteamaan identifioituvaksi takaovikriteerin perusteella, sitä voidaan yrittää etuovikriteerin avulla (teoreema 2).

Teoreema 2 *Jos joukko Z muuttujia toteuttaa seuraavat ehdot suhteessa järjestettyyn pariin muuttujia (X, Y):*

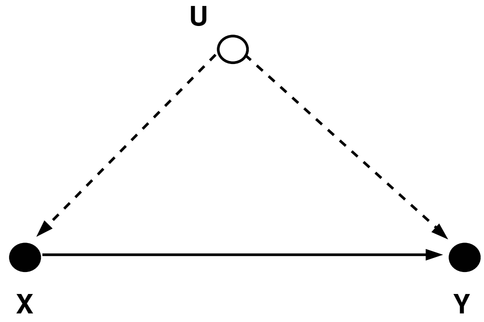
(i) Z leikkaa kaikki suunnatut polut X:stä Y:hyn,

(ii) X:n ja Z:n välillä ei ole takaovipolkua,

(iii) X d-erottaa kaikki takaovipolut Z:n ja Y:n välillä,

X:n kausaalinen seuraus Y:lle on identifioituva ja saadaan kaavasta 2.

Kuva 7:



Kausaalinen seuraus ei ole identifioituva.

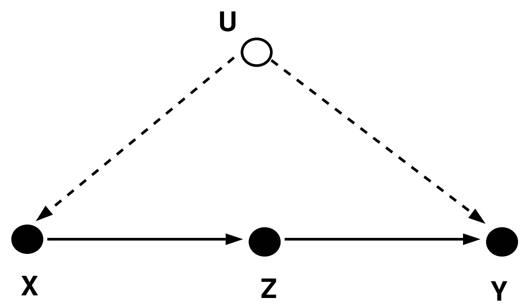
$$P(y|do(x)) = \sum_z P(z|x) \sum_{x'} P(y|x', z)P(x') \qquad (2)$$

Tarkastellaan esimerkkinä tästä havaintotutkimusta, jossa selvitetään tupakoinnin vaikutusta keuhkosyövän syntyyn. Oletetaan, että aineisto sisältää tietoja keuhkosyöpään sairastuneista ja tupakoitsijoista, ja merkitään X:llä, tupakoiko henkilö, ja Y:llä, onko hän saanut keuhkosyövän.

Ongelmana tässä tutkimuksessa on kuitenkin, että tupakoinnilla ja keuhkosyövällä saattaa olla jokin yhteinen geneettinen syy U (sekoittaja), joka voi selittää tupakoitsijoiden korkean syöpäriskin. Oletetaan lisäksi, ettei tällaisesta genotyypistä pystytä keräämään havaintoja. Kaavio 7 esittää nyt muuttujien väliset kausaaliset suhteet. Tässä mallissa X:n seurausta Y:lle ei saada identifioituvaksi kummallakaan ovikriteerillä. Takaovikriteeri toteutuisi, jos yhteisestä syystä U olisi kerätty havaintoja, mutta perimän tutkiminen voi olla hyvin hankalaa.

Etuovikriteeri saadaan kuitenkin täytettyä, jos voidaan tehdä havaintoja muuttujasta Z, joka välittää X:n vaikutuksen Y:hyn (kuva 8). Sellainen on keuhkoihin kerääntyneen tervan määrä, joka on käytännössä paljon helpommin tehtävä lisähavainto kuin geenien etsiminen. Takaoviehdot eivät puolestaan päde kaavios-

Kuva 8:



Kausaalinen seuraus on identifioituva.

sa 8, koska Z on X:n jälkeläinen ja U on tuntematon.

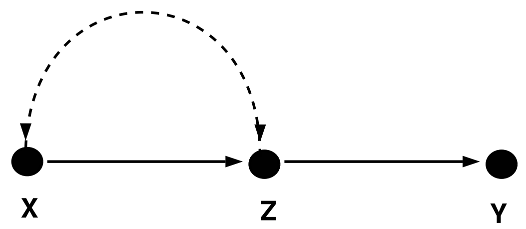
Etuovikriteeri mahdollistaa uusia hyödyllisiä tutkimusasetelmia: sekoittuminen voidaan hallita, jos pystytään tekemään havaintoja välittävästä muuttujasta Z, vaikka X:n ja Y:n takaovipoluilta ei tunnettaisikaan yhtään muuttujaa. On kuitenkin huomattava, että kaikki mahdolliset välittävät muuttujat Z on havaittava, kuten ehto (i) sanoo. Mallista ei siis saa jäädä pois yhtään suunnattua polkua X:stä Y:hyn. Jos esimerkiksi tupakointi vaikuttaa keuhkosyöpään myös jonkin muun muuttujan kuin tervan välityksellä (tai suoraan), sen jättäminen mallin ulkopuolelle aihettaa harhaa seurauksen estimointiin.

Lisäksi ehdot (ii) ja (iii) määräävät, ettei X:llä ja Z:lla ole yhteistä syytä eikä myöskään Z:lla ja Y:llä. Tämä pätee tupakointiesimerkissä: keuhkoissa olevalla tervalla tuskin on yhteistä syytä tupakoinnin tai keuhkosyövän kanssa. Puolestaan kaavioissa 9 ja 10 tällaiset syyt ovat olemassa.

2.3 Interventiokalkyyli

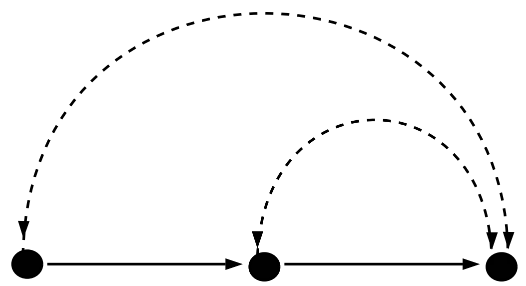
Kausaalinen seuraus voidaan myös johtaa interventiokalkyylin avulla. Kalkyylin päättelysäännöt on annettu teoreemassa 3. Olennaista sääntöjen käytössä on, että

Kuva 9:



X:n kausaalinen seuraus Y:lle ei ole identifioituva.

Kuva 10:



X:n kausaalinen seuraus Y:lle ei ole identifioituva.

mallin kaavio G määrää, milloin tiettyä sääntöä voi soveltaa. Merkinnöissä $G_{\overline{X}}$ tarkoittaa verkkoa G, josta on poistettu X:ään osoittavat nuolet. Alaviiva merkitsee, että muuttujasta lähtevät nuolet on poistettu.

Teoreema 3 (Päättelysäännöt) *Olkoon M funktionaalinen kausaalinen malli, G sen suunnatttu syklitön kaavio ja P mallin M mukainen todennäköisyysjakauma. Kaikille erillisille M:n muuttujajoukoille X, Y, Z ja W pätee:*

Sääntö 1. *Havainnon lisääminen tai poistaminen:*

$$P(y|do(x), z, w) = P(y|do(x), w), jos(Y \perp\!\!\perp Z|X, W)_{G_{\overline{X}}}. \tag{3}$$

Sääntö 2. *Teon ja havainnon vaihto:*

$$P(y|do(x), do(z), w) = P(y|do(x), z, w), jos(Y \perp\!\!\perp Z|X, W)_{G_{\overline{X}\underline{Z}}}. \tag{4}$$

Sääntö 3. *Teon lisääminen tai poistaminen:*

$$P(y|do(x), do(z), w) = P(y|do(x), w), jos(Y \perp\!\!\perp Z|X, W)_{G_{\overline{X}, \overline{Z(W)}}}. \tag{5}$$

$Z(W)$ on joukko, joka sisältää ne Z:n solmut, jotka eivät ole minkään W:n solmun esi-isiä $G_{\overline{X}}$:ssä.

Jos kausaalinen seuraus voidaan johtaa kalkyylillä muotoon, jossa on vain havaittuja muuttujia eikä lainkaan interventioita, se on identifioituva. Sääntöjen avulla voidaan johtaa esimerkiksi ovikriteerien yhteydessä esitetyt laskentakaavat.

3 Yhteenveto

Havaintotutkimuksilla voidaan tehdä luotettavia johtopäätöksiä kausaalisista seurauksista, kunhan tutkimuksen tukena on kausaalinen kaavio, joka kuvaa aineiston muutttujien väliset vaikutukset. Tämä on sangen hyödyllistä, koska satunnaisesti kontrolloitua koetta ei pystytä kaikissa tilanteissa suorittamaan. Syynä saattaa olla kokeen epäeettisyys tai korkeat kustannukset.

Kausaaliset mallit integroivat tilastollisen datan sovellusaluekohtaiseen tietämykseen. Mallin oletukset kausaalisista suhteista esitetään kaavioilla, joissa ne ovat esillä täsmällisesti ja selkeästi. Edellä esitettyjen menetelmien avulla voidaan kaaviosta päätellä, onko tutkittava seuraus laskettavissa luotettavasti havaitusta jakaumasta. Jos on, seuraus saadaan joko ovikriteerien kaavoilla tai johtamalla interventiokalkyylillä. Jos taas tunnettuja muuttujia ei ole riittävästi, menetelmä ehdottaa tarvittavia lisähavaintoja: mallista nähdään täsmällisesti sääntöjen tai ovikriteerien vaatimat havainnot. Lisäksi tutkimusta suunniteltaessa voidaan valita havaittaviksi muuttujiksi käytännössä kaikkein helpoimmin mitattavat muuttujat.

Menetelmän tulosten luotettavuus perustuu kuitenkin kaaviossa tehtyihin oletuksiin. Jos esimerkiksi jokin tuntematon yhteinen syy jää mallista pois, tulokset saattavat olla harhaisia. On siis löydettävä kaikki mahdolliset sekoittajat, mikä voi olla hyvin vaikeaa monilla sovellusalueilla. Havaintotutkimus on silloin korvattava kontrolloidulla kokeella.