David Valls

ALAN E. GELFAND

Professor d'Estadística en la Universitat de Duke (Durham, EUA)

«A LA GENT SE LI FA MOLT DIFÍCIL ENTENDRE LES ESTADÍSTIQUES»

David Conesa i Felip Pineda

«Vaig triar l'estadística simplement perquè pensava que era més interessant i tenia més vida que les matemàtiques pures, encara que tinc formació matemàtica.» Alan Gelfand explica d'aquesta manera la seua passió per l'estadística, a la qual ha dedicat la seua vida. En els noranta va popularitzar, junt amb Adrian Smith, el mètode Markov Chain Monte Carlo (MCMC), contribució que ha millorat considerablement el subconjunt de l'estadística bayesiana mitjançant el mostreig de distribucions de probabilitat. «És una d'aquelles coses amb què un es pot sentir molt afortunat perquè hi ha un munt de gent molt intel·ligent que treballa molt per a tirar endavant la seua carrera i resulta que no troben res. I un fa un descobriment i resulta ser un gran avenç. L'única cosa que puc dir en el meu favor és que almenys ho vaig aprofitar», recorda.

Durant els últims anys, la investigació d'Alan Gelfand ha girat entorn de l'estadística espaciotemporal, un camp en auge amb una gran quantitat de possibilitats, com explica humilment, però amb passió: «Hi ha una expressió que usem en anglès, *low-hanging fruit* ("fruita a l'abast de la mà"), la que es pot agafar sense gaire feina, i en l'anàlisi espaciotemporal hi havia tanta fruita a l'abast de la mà que es podia jugar i gaudir de totes les possibilitats que oferia. Moltes altres àrees s'han desenvolupat tant, s'hi ha treballat tant, que has d'arribar molt més alt per a poder trobar una mica de fruita. He tingut molta sort.» De fet, ha publicat quatre llibres i més de 250 articles científics relacionats amb aquestes qüestions, i ha rebut diversos premis. L'últim, la medalla a la distinció en la investigació de la secció d'estadística i medi ambient de l'ASA (American Statistical Association).

Actualment Alan Gelfand és professor d'Estadística en la Universitat de Duke (Durham, EUA) i membre de l'ASA, l'Institut d'Estadística Matemàtica i l'Institut Internacional d'Estadística. Es poden trobar aplicacions relacionades amb la seua investigació en àrees com l'exposició ambiental, els processos ecològics espaciotemporals i el desenvolupament de models climàtics.

«NO ÉS EL MATEIX EL TREBALL AMB 'BIG DATA' QUE L'ESTADÍSTICA. NOSALTRES INTENTEM ENTENDRE PROCESSOS COMPLEXOS, EXPLICAR, PREDIR, CAPTAR LA INCERTESA»

Quina és la seua visió sobre l'ús aplicat de l'estadística?

L'estadístic del segle XXI és un conjunt de coses. Ha de tenir un bon historial metodològic, ser bo amb els modelatges, amb la computació, amb l'anàlisi de dades, ha de poder fer visualització si cal, i té tot un conjunt d'eines que li permeten fer les coses que es necessiten avui dia. Pot donar-se el cas que diferents persones siguen més fortes en les diferents àrees, però sí que és necessari alguna cosa semblant a un paquet complet si es vol ser realment un estadístic modern. I el resultat d'això és una revitalització de l'estadística. L'estadística ha rebut una certa energia, una certa vitalitat, gràcies a tot aquest desenvolupament interdisciplinari, a tota aquesta col·laboració. I no importa gaire si es tracta de col·laborar amb les ciències mediambientals, amb la genòmica, la neurociència, les ciències socials o l'economia. En el segle XXI l'estadístic és un jugador integrat des del començament i científicament això és molt més emocionant, perquè ets verdaderament part de l'equip. Així doncs, és un moment molt emocionant per a l'estadística, en termes d'aplicacions, vull dir.

Què és exactament el pensament bayesià?

La perspectiva bayesiana és científicament intuïtiva i l'únic repte a què s'ha enfrontat el pensament bayesià vingué de dins de la comunitat estadística, no tant de la comunitat científica general. La idea que hi ha darrere del pensament bayesià és la més simple que es puga imaginar. S'infereix allò que no se sap partint d'allò que s'observa. Hi ha alguna cosa més natural que això? No obstant això la visió més clàssica està en certa manera invertida. La visió clàssica diu: què podria vostè imaginar partint del que no sap? És una manera estranya de veure les coses; un punt de vista peculiar. Jo

crec que per això el paradigma bayesià per a la inferència és tan atractiu, i la raó per la qual ens va costar tant de temps passar al capdavant de l'estadística va ser el problema computacional. Per això la comunitat bayesiana només podia resoldre problemes molt simples. Els bayesians passem molt de temps jugant amb la teoria axiomàtica formal i la perspectiva matemàtica, que va ser important en la seua fundació. Però des del punt de vista aplicat, des del punt de vista pràctic, realment no ajudava la gent.

Què va significar el mètode Markov Chain Monte Carlo (MCMC) per a l'estadística bayesiana en el seu moment?

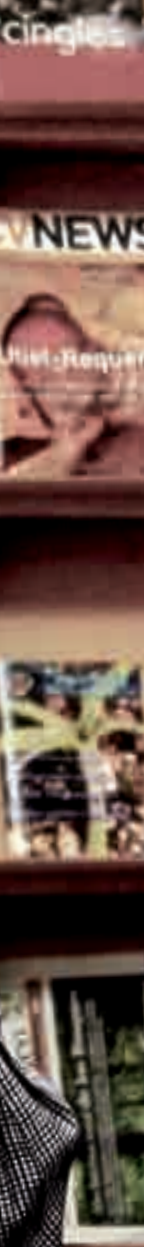
Per a un públic profà, el problema computacional bayesià té a veure amb la integració de dades d'alta dimensionalitat i altres integracions complicades que no es podien fer de manera explícita. Necessitàvem un mecanisme per a trencar el que es coneixia en aquell moment com la maledicció de la dimensionalitat, perquè els problemes d'alta dimensionalitat no es podien resoldre analíticament, i necessitàvem una estratègia. I l'estratègia resulta ser la idea més bàsica en tota l'estadística. Si es vol aprendre sobre una població, es fa un mostreig a partir d'aquesta mateixa població. Això és el que fem, i resulta que el mostreig de Gibbs i el mètode Markov Chain Monte Carlo són mecanismes que permeten la presa de mostres en poblacions complicades, en distribucions multidimensionals complicades. La veritat és que és una idea molt elegant. Però per què funciona com funciona? Perquè era impossible mostrejar una distribució de molt alta dimensionalitat, diguem-ne per exemple una distribució de cent o de mil dimensions, però sí que es pot trencar en trossos més petits i mostrejar distribucions de baixa dimensionalitat. Si ho férem de la manera correcta, amb la presa de mostres de les distribucions de baixa dimensionalitat obtindríem mostres de la distribució d'alta dimensionalitat, que és el que realment ens interessa. Aquest era el seu poder i la primera vegada que ho vam veure ens vam sentir com un xiquet amb sabates noves.

També ha estat treballant en l'estadística espacial. Quina va ser la seua experiència en aquest camp?

Em vaig adonar com era d'adequada l'estadística espacial per a la inferència bayesiana. La idea és que, per a gran part del treball estadístic, la gent usa la probabilitat i la inferència probabilística, i quan s'ajunten la incertesa i la inferència probabilística, la incertesa prové generalment de les idees asimptòtiques. I el que passa amb l'estadística espacial és que l'anàlisi asimptòtica no hi funciona, i l'única anàlisi asimptòtica que es podia usar en l'estadística espacial era la de sèries temporals, en les quals es deixava que el temps arribara fins a l'infinit. Però amb l'espai no es pot deixar que arribe fins a l'infinit. Si volem estudiar una regió, València o fins i tot Espanya, no podem ocupar-nos d'una regió de la grandària de la Terra. El que en realitat volem és pensar en una inferència en què les mostres augmenten de grandària si observem més localitzacions dins de la regió, no si fem augmentar de grandària la regió. I aquest tipus de comportament asimptòtic no funciona amb les sèries temporals clàssiques ni amb un altre tipus de teoria asimptòtica, així doncs, Bayes va molt bé per a això, perquè ens dóna la inferència exacta i

David Valls

«SI ELS ESTADÍSTICS NO SOM PROU VISIBLES, LA GENT SIMPLEMENT ASSUMIRÀ QUE NO TENIM MOLT A APORTAR»



no requereix cap argument asimptòtic. L'espai és interessant perquè quan es treballa amb sèries temporals en una dimensió hi ha l'ordre, i es pot distingir què va abans i què va després. Però en els espais no hi ha ordre. Fa que la vida siga més interessant, més exigent, més divertida, i ens ha obert moltes possibilitats. La combinació d'espai i temps és realment bella.

En quina direcció pensa vostè que progressarà l'estadística? Com troba que serà, més teòrica o més aplicada?

La contribució de l'estadística és en realitat la inferència, per tant la pregunta és si podem mantenir la inferència com a contribució científica o no. Aquí és on entren els reptes del *big data*, i la pregunta de si necessitem les estadístiques, si necessitem la inferència, o si podem operar només amb algoritmes, si podem fer ciència només amb resums descriptius i explorar bases de dades sense la necessitat de recórrer a la inferència probabilística. Abans d'això, i és interessant a causa de la crisi d'identitat a què l'estadística s'ha d'enfrontar, la pregunta és: cap on hem de dirigir-nos? I hi ha algunes persones que diran: «Bé, sempre necessitarem la part teòrica, perquè necessitem un cert rigor, assegurar-nos que la gent està fent les coses de la manera correcta.» I és cert, si ofereixes eines, la gent les utilitza. Però pot ser perillós si la gent no ha analitzat amb rigor els reptes fonamentals que part d'aquest treball comporta. Així doncs, pense que sempre hi haurà un lloc per a la part teòrica, però no hi ha dubte que el futur està en la part aplicada. I, en particular, allò que realment està canviant és l'estil de fer les coses. Ha de canviar la manera com pensem en les coses, perquè és necessari un pensament més integral, major síntesi de diferents fonts d'informació, perquè no som capaços de controlar-ho tot. Estem fonamentalment davant d'un procés més complex, amb característiques que hem d'intentar copsar de totes les maneres possibles. En resum, el futur de l'estadística consisteix a treballar en projectes interdisciplinaris complicats, a treballar en projectes i sistemes exigents.

«EL FUTUR DE L'ESTADÍSTICA CONSISTEIX A TREBALLAR EN PROJECTES INTERDISCIPLINARIS COMPLICATS, A TREBALLAR EN PROJECTES I SISTEMES EXIGENTS»

Viktor Mayer-Schönberger i Kenneth Cukier comencen el seu llibre sobre *big data*[1] explicant que les cerques en Google van predir la propagació de l'epidèmia de grip H1N1 l'any 2009. Aquest exemple serveix als autors per a citar l'article de Chris Anderson «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete»,[2] publicat el 2008 en la revista *Wired*, en el qual l'autor afirmava de manera provocadora que «els petabytes ens permeten dir: "Amb la correlació hi ha prou." Podem deixar de buscar models.» Què en pensa vostè, d'això? El futur de l'estadística passa per ser una simple anàlisi descriptiva de dades?

Antigament em considerava un estadístic, però, si em preguntes ara, em consideraria un modelador estocàstic, és a dir, modele problemes amb incertesa. La meua visió del món és que quan ens fixem en aquests processos complexos no som capaços d'explicar-los de manera perfecta i, per tant, si no podem explicar-los de manera perfecta, hem d'introduir-hi un cert grau d'incertesa, i per a mi això és el que significa el modelatge probabilístic o estocàstic. Això no ho consideraria del tot una idea del *big data*. Hi ha molts problemes que no requereixen *big data*. Pots treballar en problemes científics interessants de molta menor escala, però el que és un vertader problema en aquest camp és que l'aprenentatge automàtic i les dades massives tracten d'imaginar un gran conjunt que comprenga l'estadística, la informàtica, potser l'enginyeria informàtica i les matemàtiques aplicades. Realment aquesta no és la manera com vull pensar sobre la ciència. Crec que gran part d'aquestes coses de l'aprenentatge automàtic són una mena de cerca d'estructura en els *big data*, una cerca de patrons, de relacions. Però no és el mateix que el que fem, nosaltres intentem entendre processos complexos, explicar, intentem predir, captar la incertesa. No és la mateixa manera de veure les coses, no pot respondre a les mateixes preguntes. Necessitem ciència amb dades, és clar, i la ciència no és ciència sense estadístiques. Realment crec que l'enfocament de la ciència que estic descrivint és molt diferent a una simple exploració d'una base de dades enormes per tractar de traure'n alguna estructura, que és el més comú en l'aprenentatge automàtic.

Ens agradaria parlar de la visibilitat de l'estadística. Què han de fer els estadístics per a fer-la més visible?

[1] Mayer-Schönberger, V. i K. Cukier, 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. Eamon Dolan/Houghton Mifflin Harcourt. Boston.

[2] Anderson, C., 2008. «The End of Theory: The Data Deluge Makes the Scientific Method Obsolete». *Wired*, 16 (7). Disponible en: <http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory>.

L'any passat va ser l'Any Internacional de l'Estadística, una oportunitat perquè pensàrem un poc més en com continuarem participant en la comunitat científica més gran. Ens va anar bé. Històricament, els estadístics s'han acontentat d'estar en un segon pla. Hem estat gent de baixa visibilitat. No necessitàvem eixir en *El País* o en *The New York Times*. No necessitàvem eixir per televisió. És a dir, el líder científic del projecte solia ser un estadístic a qui no li importava estar en segon pla. Crec que això té moltes coses bones, en el sentit que fa que la comunitat estadística siga més amistosa, un poc menys cruel, un poc més solidària. Perquè no busquem tant la glòria, no competim tant, ens donem més suport, i aquesta mena d'innocència és en veritat una cosa molt bona. No vull perdre això, però reconec que, si no som prou visibles, la gent simplement assumirà que no tenim molt a aportar. Es pensaran que un científic d'una altra especialitat pot fer el que fem nosaltres i que no tenim un paper molt important que exercir. Realment, no sé com evitar aquest problema, perquè no vull sacrificar la innocència, però d'altra banda tampoc vull estar fora en cerca de la glòria. No és una motivació. Per a mi l'important és la ciència. Exercir bona ciència, tractar de resoldre els problemes importants, i tractar de fer una contribució als problemes importants. Pense que aquest és el repte en realitat.


David Valls

«HI HA UN VERTADER REPTE EN EL MAL ÚS DE LES DADES I AIXÒ CREA ESCEPTICISME EN LA MENT DEL PÚBLIC, CREA UN VERTADER DUBTE SOBRE LA VALIDESA DE L'ANÀLISI ESTADÍSTICA»

Quina és la seua opinió sobre el futur del finançament en la investigació?

Fa temps podies proposar un programa d'investigació que fóra estadística teòrica pura i dura i, bé, potser no et feies ric, però podies obtenir una mica de diners per finançar-la i potser et donaven diners per a disposar d'un o dos estudiants, per a uns pocs viatges, i aquesta era més o menys la manera de fer les coses. Però ara, aquesta mena d'investigació no serà fàcil de finançar, almenys als Estats Units –i sé que al Regne Unit, i potser a Espanya, tampoc–, com que la crisi ha dificultat tant les coses, aconseguir finançament és més complicat, però sé que el model que els organismes finançadors busquen és una idea de projecte interdisciplinari. Pense que segurament aquest serà el futur del finançament. La dificultat és que la quantitat de fons s'està reduint perquè la gent aliena a la comunitat científica no valora gaire la ciència bàsica, pensa que el govern no hauria de gastar diners en ciència quan hi ha altres coses en què hauria de gastar-ne, i potser que no hauria d'haver gastat tants diners en primer lloc. I així, el que està succeint és que la quantitat de finançament s'està reduint però hi ha més pressió perquè, almenys als Estats Units, les universitats esperen que el seu professorat els reporte diners. Així doncs, és una partida que no poden guanyar. No sé com anirà. Algunes persones suggereixen que la indústria privada ajude. Jo no estic tan segur, trobe que les indústries privades tenen les seues pròpies prioritats. No pense que puguem esperar que proporcionen el finançament que vulguem per a mantenir la investigació científica bàsica. No estic segur que hi haja una solució.

El mal ús de l'estadística en els mitjans de comunicació pot donar lloc, a propòsit o sense voler, a la manipulació de les xifres. És això un problema comú al seu parer?

La vella mentida solia ser, almenys en anglès, *liars, damned liars and statisticians* ("mentiders, maleïts mentiders i estadístics"). La idea era que els estadístics podien abusar de les dades. No hi ha dubte que s'ha fet. El que tenim ara és una gran pressió per a publicar, i

aquesta gran pressió per a publicar comporta una pressió per trobar coses. Aquesta mena de pressió condueix a un mal ús o una tergiversació dels resultats estadístics. Això s'uneix al repte que anomenem reproductibilitat. Per exemple: si a un altre equip científic li dones un conjunt de dades sobre les quals has escrit alguna cosa, aquest equip podrà reproduir allò que tu has esbrinat? I la història d'això és notablement decebedora, depèn del camp, però fa poc vaig escoltar una meravellosa xarrada sobre la reproductibilitat d'estadístiques mèdiques, i els nombres espanten. S'ha pogut reproduir menys del 50 % dels conjunts de dades que s'han publicat. Hi ha un veritable repte en el mal ús de les dades i això crea escepticisme en la ment del públic, crea un vertader dubte sobre la validesa de l'anàlisi estadística, i la sensació entre el públic és que un estadístic pot explicar-te la història que vulga perquè pot manipular les dades al seu capritx. I en un cert sentit és veritat.

Considera vostè que la població està preparada per a entendre correctament l'anàlisi estadística en els mitjans de comunicació?

Estic profundament preocupat per les nocions numèriques, com en diem nosaltres, de la població. No em sembla que molta gent entenga ni tan sols els ordres de magnitud. Sé, per exemple, que si entre en una botigueta dels Estats Units, compre una barreta de xocolata i done vint dòlars, la persona de darrere de la màquina registradora no sap quants diners m'ha de tornar. No tenen ni idea d'aritmètica. L'única manera que tenen de saber-ho és introduir el número i esperar que la màquina els diga el canvi. I sé que això no passa només als Estats Units, sinó pertot arreu. Pense que hi ha una pèrdua completa de la magnitud dels nombres, per això a la gent se li fa molt difícil entendre les estadístiques. A més de tot això, quan proporcionem relacions estadístiques, encara que la majoria de la gent siga capaç d'entendre les mitjanes, potser les proporcions, com els percentatges, quan intentes anar més enllà, fins a la incertesa o la variabilitat, la gent té moltes dificultats amb aquestes qüestions, fins i tot els que tenen una bona formació. Totes aquestes coses dificulten molt que el públic accepte i comprenga les estadístiques a gran escala. Al meu parer hem malacostumat la gent a no apreciar els nombres perquè tots podem fer càlculs senzills en un mòbil i tot està automatitzat quan comprem, sense que hàgem de fer res.

David Conesa. Professor del departament d'Estadística i Investigació Operativa de la Universitat de València.
Felip Pineda. Periodista. Revista *Mètode*, Universitat de València.