Principis d'inferència bayesiana: la revolució a posteriori

Escrit per
Ferran Palero Pastor
Departament de Genètica de la Universitat de Barcelona

L'estadística bayesiana ha revolucionat la genètica moderna, aportant una nova eina d'anàlisi i revifant àrees d'estudi que havien quedat estancades des de feia dècades. Gràcies als mètodes bayesians, podem treballar amb models evolutius molt més complexos. Una de les àrees on més s'ha fet notar la revolució bayesiana és a la inferència filogenètica. Però no és l'única; a la genètica de poblacions els mètodes bayesians també s'han fet servir per a estimar paràmetres en diferents models demogràfics. Això ens permet assignar individus a les seues poblacions originals, estimar canvis en el volum de les poblacions i detectar l'acció de la selecció sobre els gens. A escala genòmica, l'estadística bayesiana s'ha fet servir per a realitzar inferències sobre els nivells d'expressió gènica i per a la localització de gens al genoma.

Els humans sempre intentem trobar explicacions per tot allò que ens envolta. Les explicacions més simples es basen en causes directes, en models deterministes. Tanmateix, en el món real ens trobem amb casos complexos que no podem explicar de forma directa. Aquests tipus de casos (els més habituals), en els quals el valor resultant d'una gran quantitat d'interaccions no es pot predir amb certesa, s'expliquen mitjançant models probabilístics. Un model probabilístic especifica una distribució de probabilitats conjunta per a les variables aleatòries. Tant les dades com els paràmetres del model són variables aleatòries, donat que no podem predir els seus valors amb certesa.

En l'estadística clàssica, la probabilitat d'un succés s'interpreta com la proporció dels casos en què es realitza el succés; per posar l'exemple més conegut, el nombre de cares obtingudes quan llancem una moneda x vegades. Aquesta probabilitat es considera un valor absolut. Tanmateix, en inferència bayesiana, les probabilitats no s'interpreten com a freqüències o proporcions, sinó més aviat com a graus de certesa. És evident que mai podrem arribar a la veritat absoluta sobre el món real, però conforme acumulem evidències el nivell de credibilitat d'una hipòtesi varia. En definitiva, cal tenir en compte que un conjunt d'observacions suposa un mitjà per a canviar el nivell de credibilitat de la nostra hipòtesi, no una forma d'aconseguir la veritat absoluta.

L'estadística clàssica es limita a rebutjar una hipòtesi quan creu que les dades no responen a allò que calia esperar sota unes condicions determinades. Per altra banda, la inferència bayesiana fa servir una estimació del grau de credibilitat d'una hipòtesi abans d'observar noves dades per proporcionar-nos el nivell de credibilitat d'eixa hipòtesi després d'observar les dades. Els bayesians no creuen escaient especificar una hipòtesi i decidir simplement si rebutjar-la o no rebutjar-la. Per ells, les lleis de probabilitat han d'indicar-nos quines hipòtesis hem de creure i fins a quin nivell podem confiar-hi, sense descartar-ne cap.

Principis bàsics

El teorema de Bayes és un resultat de la teoria de probabilitats que ens proporciona la distribució de probabilitats *condicionada* d'una variable aleatòria [P (A | B)]. Devem aquest teorema al reverend Thomas Bayes (1702-1761) (**Fig. 1**), que va morir abans de fer coneguda la seua teoria de la probabilitat inversa. Tot i això, avui en dia encara se li reconeix el mèrit d'haver proposat aquesta idea.

Figura 1. El matemàtic britànic Thomas Bayes (1702-1761) va estudiar a la Universitat d'Edinburgh, ja que Òxford i Cambridge li tenien les portes tancades per ser un no-conformista (*nonconformist*). Va ser ordenat ministre presbiterià com ho havia estat el seu pare, Joshua Bayes, un dels sis primers no-conformistes a ser ordenats a Anglaterra. El 1742 va ser escollit com a membre (*Fellow*) de la Reial Societat, segurament per la força del seu treball *Introduction to the Doctrine of fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst*, publicat anònimament el 1736.

© Ferran Palero

L'anàlisi filogenètica de les llagostes també ha permès descobrir que aquesta larva (*Scyllarus pygmaeus*) havia estat mal identificada des de feia més de cent anys.

Per obtenir eixa distribució de probabilitats cal:

1. Informació d'una variable aleatòria B, com a distribució condicionada de B, donat A [P (B|A)].

2. La distribució marginal d'A [P (A)].

El teorema de Bayes simplement relaciona les probabilitats condicionals i marginals:

P (A | B) P (B) = P (A, B) = P (B | A) P (A)

Sent P (A, B) la probabilitat conjunta d'A i B.

L'anomenat teorema de Bayes no va ser mai especificat per Bayes en la seua forma més habitual:

P (Ho | D) = P (Ho) * P (D | Ho) / P (D)

El teorema diu que la distribució posterior [P (Ho|D)], on considerem l'evidència a favor d'uns valors paramètrics determinats, es pot estudiar mitjançant la distribució de probabilitats que assignem a priori P (Ho), la funció de versemblança P (D | Ho) i una constant normalitzadora P (D).

Cal tenir en compte que és molt important assignar correctament les probabilitats a priori que donem a la nostra hipòtesi i a les alternatives, i a més, calcular les probabilitats condicionals de les dades per a cada hipòtesi.

La **distribució de probabilitats a priori [P (Ho)]** inclou la informació sobre els valors d'un paràmetre abans d'examinar les dades, en forma d'una distribució de probabilitats. És la distribució de probabilitats del paràmetre obtinguda a partir dels nostres coneixements previs.

La **versemblança o *likelihood* [P (D | Ho) ó L (Ho | D)]** és una distribució condicional que especifica la probabilitat de les dades observades donat qualsevol valor particular dels paràmetres. S'obté a partir d'un model estadístic del procés estudiat en el qual és necessari i fonamental considerar *com expliquen les dades* els paràmetres.

La **distribució conjunta [P (Ho) * P (D | Ho)]** representa la distribució de probabilitats de totes les combinacions possibles de dues o més variables aleatòries. És el producte de la versemblança i de la distribució a priori, donat que ambdues funcions combinen tota la informació disponible sobre els paràmetres.

El principal objectiu de la inferència bayesiana és calcular la distribució posterior dels paràmetres, donades les dades (**Fig. 2**). La inferència bayesiana suposa, doncs, la manipulació de la distribució conjunta per tal de fer inferències al voltant dels paràmetres, o del model de probabilitats, donades les dades.

Així, les **distribucions marginals** s'obtenen integrant la distribució conjunta sobre les dades (de forma que obtenim la distribució a priori [P (Ho)]) o sobre els valors paramètrics (a través dels quals obtenim la versemblança marginal o distribució predictiva a priori [P (D)]).

Les **distribucions condicionals** s'indiquen amb les línees puntejades a la figura 2. Representen com fer un «tall» a través de la distribució conjunta i refer l'escala de la distribució, de forma que la suma (integral) de totes les dades possibles siga igual a la unitat. Aquest factor d'escala ens el proporciona la distribució marginal: qualsevol distribució condicional és simplement la distribució conjunta dividida per una distribució marginal.

La **distribució posterior**, que és la que ens interessa en inferència bayesiana, s'obté dividint la distribució conjunta per la versemblança marginal. Obtenir la versemblança marginal és, doncs, el que resulta típicament problemàtic. La versemblança marginal és la distribució de probabilitats de les dades independentment dels valors paramètrics. És una constant normalitzadora que es pot calcular com la suma de totes les hipòtesis mútuament excloents.

Mètodes amb cadenes de Markov-Monte Carlo (MCMC)

Una cadena de Markov és un model matemàtic que podem fer servir per a modelar una seqüència de variables aleatòries. En eixa seqüència, la probabilitat que una variable tinga un valor determinat depèn tan sols del valor d'un nombre determinat de variables precedents (*propietat de Markov*). Així, en una cadena d'ordre *n*, la distribució de probabilitats d'una variable depèn de les *n* observacions anteriors.

En determinades condicions, una cadena de Markov pot tindre una distribució estacionària. Això vol dir que els estats visitats tendeixen a una distribució de probabilitats específica que no depèn del nombre de repeticions, ni de l'estat inicial de la variable. En aquests casos, el que desitgem és construir una distribució estacionària que siga com la distribució que ens interessa i mostrejar-la per a realitzar inferències. En l'anàlisi bayesiana, aquesta distribució d'interès és la distribució de probabilitats posteriors conjunta per a un o més paràmetres.

La idea bàsica de la integració per Monte Carlo és que les propietats de les variables aleatòries (com per exemple, la mitjana) poden estudiar-se simulant molts casos d'una variable i analitzant els resultats. Cada rèplica de les simulacions per Monte Carlo és independent, així que el procés és equivalent a prendre repetides mostres d'una cadena de Markov estacionària en punts suficientment separats com perquè no estiguen correlacionats.

El mètode Monte Carlo té l'avantatge que les estimacions no són biaixades i que l'error estàndard de les estimacions es pot valorar, perquè les variables aleatòries simulades són independentment i igual distribuïdes. Tanmateix, cal tenir en compte que és molt difícil elaborar un mostrejador MCMC adequat. Aquest pot donar passes llargues o passes curtes, i reconèixer si la cadena és prou llarga o quan ha arribat a la convergència no és gens fàcil. Els mètodes MCMC també s'han fet servir per a estimar la versemblança en inferència per *maximum likelihood*.

Els mètodes MCMC i el seu ús en les filogènies moleculars

Una filogènia és una representació, normalment en forma d'arbre filogenètic, de les relacions de parentiu entre grups taxonòmics. Per tal de reconstruir la història evolutiva d'un grup taxonòmic a partir de dades moleculars, necessitem un model d'evolució molecular que ens explique com canvien les molècules al llarg del temps. En filogenètica, l'anàlisi bayesià suposa especificar un model (*likelihood*) i una distribució a priori, i aleshores integrar el producte d'aquestes quantitats sobre tots els valors paramètrics possibles per a determinar la probabilitat posterior de cada arbre. Donada la gran quantitat de paràmetres

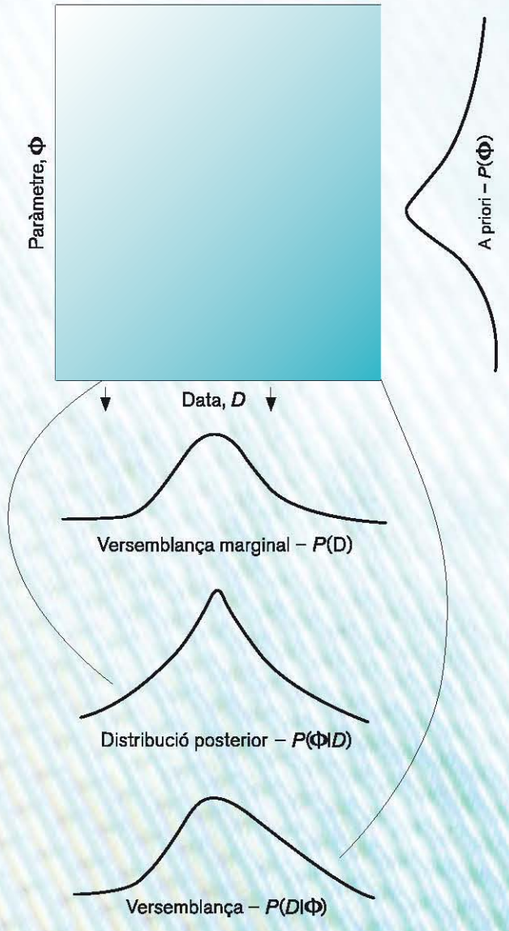

Figura 2. La inferència bayesiana suposa la manipulació de la distribució conjunta per tal de fer inferències al voltant dels paràmetres, o del model de probabilitats, donades les dades.

© Poseidon Ocean Sciences (Nova York, EUA).

Un dels molts *puerulus* (*Panulirus argus*) del centre de cultiu de Poseidon Ocean Sciences, a Florida.

que contenen, les funcions de versemblança per als models filogenètics són massa complexes com per ser integrades analíticament. Així que les aproximacions bayesianes es recolzen en mètodes Markov-Monte Carlo.

Com hem comentat anteriorment, MCMC genera una distribució de probabilitats seguint una sèrie de passos encadenats. El pas següent s'obté mitjançant l'alteració aleatòria d'alguns dels paràmetres del model. Si el pas que hem donat ens proporciona una densitat de probabilitat posterior més elevada, acceptem el moviment, si no, normalment serà rebutjat (no sempre, alguns passos enrere xicotets són acceptats!). La cadena de passos creada representa un desplaçament pels diferents conjunts d'arbres i models evolutius. La cadena tendirà a romandre en regions amb una probabilitat posterior elevada.

La localització en l'espai paramètric suposa una descripció de l'arbre i una especificació del model evolutiu. Per definició, la proporció del temps que una cadena passa en una regió concreta de l'espai paramètric es pot fer servir com a estimació de la probabilitat posterior d'eixa regió. Al final de l'anàlisi, tenim una estimació de la probabilitat de l'arbre donades les dades, que és el que ens interessa. Per suposat, aquesta estimació es recolza en un model evolutiu concret i en què les distribucions a priori siguen raonables.

Així, al Departament de Genètica de la Universitat de Barcelona, estudiem les relacions filogenètiques entre distintes espècies de llagosta marina de tots els oceans. Gràcies a la seqüenciació i estudi de diferents fragments del genoma d'aquests crustacis mitjançant les eines que ens proporciona l'estadística bayesiana, hem estat capaços de recuperar informació sobre esdeveniments que van ocórrer fa més de 200 milions d'anys..., fins i tot abans que aparegueren els dinosaures! I

Referències

Principis bàsics

Berger, J. O. (1985). *Statistical Decision Theory and BayesianAnalysis* (2a ed.). Springer-Verlag.

Berry, D. A. (1996). *Statistics: A Bayesian Perspective*. Wadsworth Publishing.

Gelman, A. [*et al.*] (1995). *Bayesian Data Analysis*. Chapman & Hall.

Mètodes amb cadenes de Markov-Monte Carlo

Gamerman, D. (1997). *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman & Hall.

Hastings, W. K. (1970). «Monte Carlo sampling methods using Markov chains and their application». *Biometrika*, 57, 97–109.

Huelsenbeck, J. P. [*et al.*] (2002). «Potential applications and pitfalls of Bayesian inference of phylogeny». *Syst. Biol.*, 51, 673–688.

Lewis, P. O. (2001). «Phylogenetic systematics turns over a new leaf». *Trends Ecol. Evol.*, 16, 30–37.

Metropolis, N. [*et al.*] (1953). «Equations of state calculations by fast computing machine». *J. Chem. Phys.*, 21, 1087–1091.

Ripley, B. D. (1987). *Stochastic Simulation*. Nova York: Wiley and Sons.

Ross, S. M. (1997). *Simulation*. Nova York: Academic.

Ferran Palero (Alzira, 1981)

És llicenciat en ciències biològiques per la Universitat de València (2004) i va obtenir el DEA en Genètica el 2006. Des del 2005 realitza la tesi doctoral sota la direcció de Marta Pascual (Departament de Genètica, UB), Pere Abelló (ICM-CSIC) i Enrique Macpherson (CEAB-CSIC). El seu projecte de tesi es titula *Genètica de les poblacions de llagosta vermella* (Palinurus elephas) *de l'Atlàntic i el Mediterrani*, i forma part del projecte *Genética evolutiva de procesos colonizadores y análisis molecular de la biodiversidad*, dirigit pel Dr. Lluís Serra Camó (Departament de Genètica, UB).