HAL
archives-ouvertes.fr

Changement de point de vue : de l'interprétation de données à la modélisation stochastique grâce aux approches bayésiennes.

Jean-Baptiste Denis

▶ **To cite this version:**

Jean-Baptiste Denis. Changement de point de vue : de l'interprétation de données à la modélisation stochastique grâce aux approches bayésiennes.. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. 2009. <inria-00386622>

HAL Id: inria-00386622
https://hal.inria.fr/inria-00386622

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Changement de point de vue : de l'interprétation de données à la modélisation stochastique grâce aux approches bayésiennes.

Jean-Baptiste Denis

Unité de recherche MIA INRA F-78352 Jouy-en-Josas

Abstract

Statisticien plongé dans un domaine d'application où les données sont rares et hétérogènes, l'auteur a trouvé recours dans l'usage de deux approches complémentaires : la construction de réseaux bayésiens comme base de la modélisation et la pratique de la statistique bayésienne pour l'interprétation des données. Il relate son expérience en tentant de la présenter de manière générique. Dans une première partie, les deux approches sont brièvement rappelées pour montrer qu'à part le théorème de Bayes, elles n'ont rien en commun. S'appuyant ensuite sur la modélisation du nombre de campylobactérioses en France liées à la consommation de poulets, un nouveau point de vue est suggéré pour l'interprétation de données. Il s'agit (1) de modéliser en soi le phénomène d'intérêt à l'aide d'un réseau bayésien ; puis (2) de l'étendre pour définir la vraisemblance des données disponibles et intégrer l'information qu'elles contiennent par conditionnement, c'est à dire en appliquant le principe de la statistique bayésienne. R et les logiciels de la famille Bugs se sont révélés bien adaptés pour la réalisation pratique de cette proposition.

Applied statistician involved in a field where data are rare and heterogeneous, the author found help from two complementary but independent Bayesian approaches : the setup of a Bayesian network to model the phenomemum under study and the use of a Bayesian statistical procedure to get information from the available data sets. This global and generic approach is presented through the modelling of the number of campylobacterioses in France due to the consumption of broilers. Contrary to standard approaches of Monte Carlo simulations, data sets can be used at every point of the chain, even downstream of other data sets. R and softwares of the Bugs family are well adapted to perform the proposed procedure.

Analyse de données, Modélisation, Réseau bayésien, Statistique bayésienne.

1 Introduction

Statisticien appliqué en recherche agronomique depuis une trentaine d'années, j'ai pu observer une évolution notable dans mes collaborations avec les chercheurs biologistes. Celle-ci n'a pris sa véritable ampleur que par l'utilisation de deux approches bayésiennes

dont je n'ai bien séparé les concepts qu'après quelques années d'approfondissement. Ce sont ces idées que je voudrais exposer dans ce papier.

Au départ, travaillant principalement en génétique végétale, les interprétations statistiques que je conduisais étaient centrées sur les données. Le modèle probabiliste utilisé était surtout justifié par la manière dont les données avaient été recueillies. S'agissait-il d'une enquête ou d'une expérimentation ? Comment avait été déterminés les niveaux des facteurs étudiés ? Quelles étaient les combinaisons des données disponibles ? Tels étaient les critères clef qui permettaient de proposer ou pas telle analyse statistique, de retenir ou pas tel terme d'interaction dans le modèle. Le degré le plus achevé était la prise en compte correcte des *randomisations* des plans d'expériences conduisant à des analyses distinctes pour les split-plots, les carrés latins et autres blocs incomplets...

Plus tard, il m'a fallu m'investir dans le domaine de l'appréciation quantitative des risques microbiologiques liés à l'alimentation. Les données n'étaient plus ordonnées en matrices serrées, elles étaient quasi inexistantes ! Quand on en trouve, elles proviennent de différentes publications scientifiques réalisées pour répondre à d'autres questions, sinon on en est réduit à consulter les experts qui, au fil de leur expérience, en fonction de leur personnalité, se sont créés des références implicites et qualitatives leur permettant d'émettre des avis dans certaines situations concrètes. C'est pour avancer dans ma fonction de *quantitativiseur* que je me suis peu à peu tourné vers les approches bayésiennes. Elles ont permis, me semble-t-il, de relever le challenge.

Dans ce papier est établie la distinction entre réseaux bayésiens et statistique bayésienne (en fait leur qualificatif commun ne repose que sur l'usage central du théorème des probabilités conditionnelles) et la proposition de les conjuguer.

2 Réseaux bayésiens et Statistique bayésienne

2.1 Réseaux bayésiens

Les réseaux bayésiens sont une façon commode de définir la loi de probabilité conjointe d'un ensemble de variables aléatoires au moyen de conditionnements successifs. Leur intérêt réside dans la possibilité de proposer facilement des modèles et sous-modèles par simplifications de ces conditionnements. La loi conjointe, $[A, B, C, D, E]$, des cinq variables aléatoires $\{A, B, C, D, E\}$ peut toujours s'écrire comme le produit de $[A]$, $[B \mid A]$, $[C \mid A, B]$, $[D \mid A, B, C]$ et de $[E \mid A, B, C, D]$ où la barre verticale signale le conditionnement. Si cette forme générale peut être simplifiée en

$$[A, B, C, D, E] = [A]\,[B \mid A]\,[C \mid B]\,[D \mid B]\,[E \mid C, D] \tag{1}$$

on peut vérifier, sans avoir précisé leurs distributions, qu'un certain nombre d'indépendances conditionnelles s'en déduisent. Ici :

$$[E \mid A, B] \;=\; [E \mid B]$$

Figure 1: Graphe acyclique dirigé du réseau bayésien de l'équation (1).

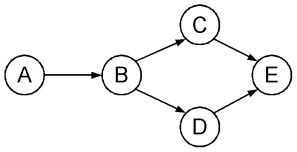
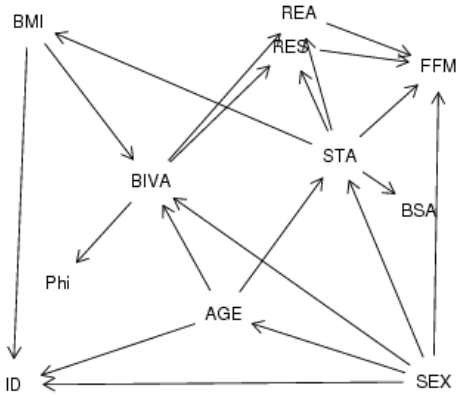


Figure 2: Composition corporelle (avec L. Mioche (INRA)).



$$C \mid B \quad \perp \quad D \mid B$$

En fait l'attractivité des réseaux bayésiens est de visualiser ces propriétés au moyen d'un graphe acyclique dirigé dont les noeuds sont les variables aléatoires et les variables conditionnantes de chacune sont les parents (figure 1).

Dans la pratique, on procède à l'inverse : on bâtit le graphe des relations directes entre les variables qu'on veut/doit considérer, le plus souvent en se basant sur les causalités qu'elles entretiennent, puis on définit pour chacune une distribution conditionnelle (ou marginale). Le résultat est qu'on obtient ainsi de manière simple car locale (pour chacune des variables) une modélisation stochastique complexe, globale et parfaitement cohérente sur l'ensemble des variables. Ajouter ou retrancher une variable d'un réseau bayésien s'opère également de manière locale, tout en conservant la jointe marginale des autres variables.

Les réseaux bayésiens sont un outil commode pour définir des probabilités parcimonieuses en paramètres et fidèles à une vision de la réalité. Ils se révèlent extrêmement efficaces pour envisager une modélisation avec des personnes rebutées par les formalisations mathématiques ; ce sont des supports utiles pour la réflexion dans le cas de modèles complexes (figures 2 et 4). A relever que les réseaux bayésiens ont principalement été étudiés en Intelligence Artificielle et assez peu en Statistique.

Figure 3: Statistique bayésienne sous forme de réseaux bayésiens.

2.2 Statistique bayésienne

Soit un vecteur d'observations y que l'on veut interprêter. Pour cela on suppose qu'il est la réalisation d'une variable aléatoire, Y, dont la distribution, dite vraisemblance, dépend d'un certain nombre de paramètres notés par un vecteur Θ. En statistique classique, on admet que Θ prend une valeur fixe mais inconnue, disons θ, que le statisticien doit s'efforcer d'approcher au mieux. En statistique bayésienne, Θ est considéré comme une variable aléatoire et il faut en préciser la loi, dite priore, pour faire, grâce au théorème de Bayes, une inférence à partir de y. La démarche utilisée s'illustre parfaitement à l'aide du réseau bayésien à deux noeuds de la figure (3.a). L'inférence n'est autre que le retournement de l'arc du réseau bayésien (ce qui est toujours possible ; figure (3.b)) produisant la postériore : $[\Theta \mid Y = y]$.

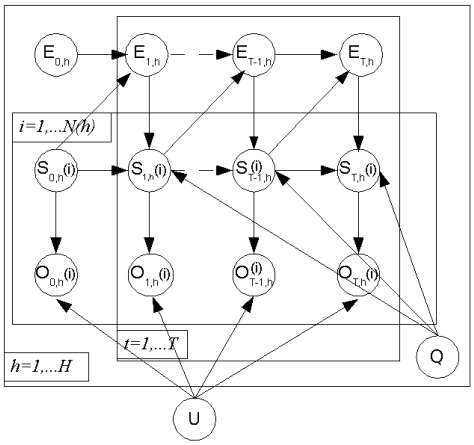
2.3 Réseaux bayésiens et statistique bayésienne

Tels que présentés, les réseaux bayésien et la statistique bayésienne n'ont rien à voir : définir un modèle probabiliste ou extraire de l'information d'un ensemble de données. Dans les pratiques diverses que l'on peut rencontrer les choses ne sont pas aussi tranchées. Par exemple, beaucoup d'utilisations des réseaux bayésiens consistent à conditionner les variables cibles par l'*instanciation* (ou *évidenciation*) d'une ou de plusieurs autres variables. Les variables conditionnantes ne sont pas considérées comme des données, parfois il ne s'agit que de raisonnement "what if" : quel effet aurait l'occurrence de telle valeur pour telle variable ? D'autre part, certains logiciels d'analyse statistitique bayésienne se servent - sans les nommer - de réseaux bayésiens pour définir les priores et les vraisemblances, c'est aussi le cas en statistique classique pour la définition de modèles hiérarchiques où les paramètres impliqués dans la définition de la vraisemblance sont des enfants d'hyperparamètres plus globaux.

3 Modélisation stochastique appuyée par la statistique bayésienne

Lors de la modélisation du nombre de campylobactérioses en France conséquence de la consommation de poulets [a] nous avons suivi une stratégie alliant les deux approches bayésiennes décrites. Dans un premier temps la modélisation du système d'intérêt est

Figure 4: Evolution d'un épidémie (avec A. Courcoul (ENVN) et E. Vergu (INRA)).



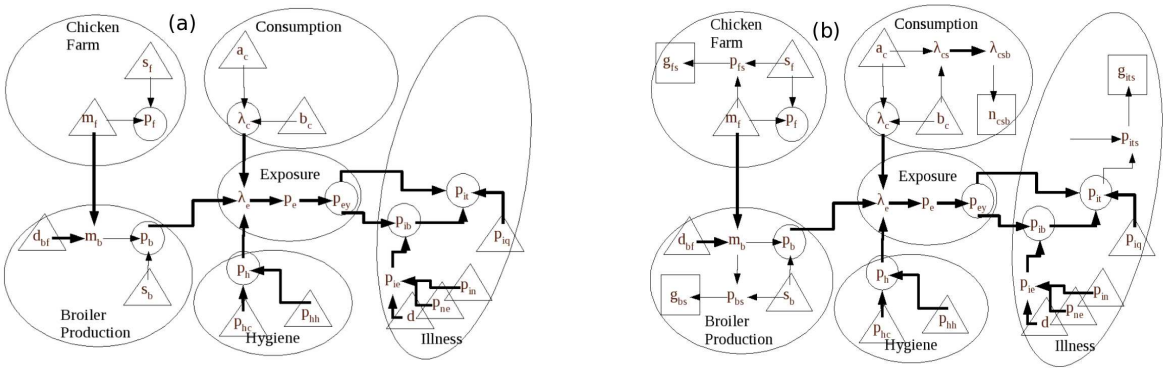
menée par la construction d'un réseau bayésien, puis une inférence statistique bayésienne est ensuite réalisée pour y incorporer les rares jeux de données disponibles. Cette démarche s'est avérée utilisable et utile dans d'autres situations (figures 2 et 4). Elle est esquissée ici sur l'exemple campylobacter-poulet.

Les campylobacters sont des bactéries très fréquentes dans l'intestin des volatiles. Les carcasses de poulet peuvent être contaminées et provoquer, par l'intermédiaire d'un aliment non cuit comme une salade, une gastro-entérite, en général sans grave conséquence. Mais si l'impact individuel est faible, le coût économique supporté par la société du fait des arrêts de travail en fait une des maladies liée à l'alimentation des plus importantes.

L'approche proposée se décline en un certain nombre d'étapes simples successives, agrémentées de nombreux aller-retours : (1.a) listage et définition des variables du système qui seront les noeuds du réseau bayésien, (1.b) établissement des relations directes entre les noeuds sur lesquelles sont élaborées (1.c) les distributions conditionnelles, (1.d) génération de pseudo-données pour apprécier la cohérence globale de la construction, (2.a) extension du réseau bayésien pour inclure les données disponibles, (2.b) conditionnement du réseau bayésien par les valeurs observées, (2.c) génération de pseudo-données pour apprécier la cohérence globale de la construction et par comparaison avec les résultats obtenus en (1.d) juger de l'information apportée par les données. Du point de vue de la statistique bayésienne, les étapes (1) correspondent à la définition des paramètres et de la priore portée sur eux, les étapes (2) à la définition de la vraisemblance et au calcul de la postériore.

La figure 5.a présente la structure du modèle initial des étapes (1). Sans entrer dans les détails, on constate que le processus est modélisé en grands modules (la contamination par la bactérie est suivie de la ferme à la cuisine en passant par l'abattoir ; l'occurrence de la maladie dépend de la consommation et de la dose-réponse choisie). Cette modélisation est fort ambitieuse, elle rassemble dans une même formalisation des connaissances zootech-

Figure 5: (a) Occurence de campylobactérioses dues à la consommation de poulets ; (b) modèle complété pour intégrer les données disponibles (noeuds carrés).



nique, agro-industrielle, comportementales, économiques, médicales, épidémiologiques... mais elle est aussi rudimentaire puisque chacune est résumée par quelques variables aléatoires. Elle présente cependant l'avantage de proposer un formalisme unique conduisant les spécialistes impliqués vers un référentiel commun. L'élicitation des priores par les experts est à la fois fondamentale et difficile. L'usage du graphe du modèle et des simulations de variables observables sont deux appuis majeurs.

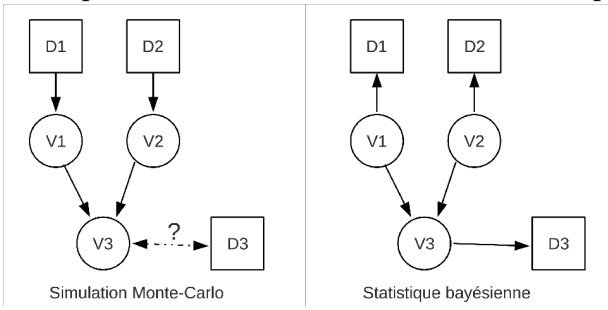
On comprendra dans cette perpective d'utilisation de la statistique bayésienne que l'optique n'est pas de rechercher des priores les moins informatives possibles, mais au contraire les priores qui soient les plus fidèles aux idées des experts. En particulier, si la démarche proposée libère de la contrainte d'identifiabilité des paramètres par les données, c'est qu'elle repose sur l'information apportée par la priore.

La figure 5.b présente la structure du modèle complété pour y intégrer les données. Il apparaît, comme on pouvait s'y attendre, que la définition du modèle initial n'est pas complètement indépendante des données disponibles. On peut quand même souligner la difficulté que soulève leur interprétation globale sans optique bayésienne. La pratique classique est en effet de faire indépendamment l'estimation locale des paramètres (basée sur les connaissnces d'expert et les données disponibles) des différents modules, puis de les utiliser pour simuler le phénomène... avec un gros hic lorsque des données se trouvent en aval de données amont (figure 6) .

4 Logiciels utiles

Dans la plupart des cas, la démarche proposée peut se réaliser avec les logiciels de la famille Bugs [e, b]. Utilisant une syntaxe proche du codage d'une programmation R [d], ils autorisent le codage de modèles très complexes en peu de lignes. La variété des distributions de probabilité disponibles est grande, elle inclut un certain nombre de

Figure 6: Comparaison des démarches habituelle et proposée.



distributions multivariables continues (Normale, Student, Wishart, Dirichlet) et discrète (multinomiale). Les algorithmes de simulation MCMC qu'ils mettent en oeuvre sont, comme il est écrit en rouge en première page de leur aide, délicats à manier et l'usager doit connaître un minimum du fonctionnement des échantillonneurs. Mention doit être faite de la possibilité d'utiliser ces logiciels depuis le langage R grâce à des paquets adaptés pour définir les entrées ou pour traiter les échantillons MCMC obtenus.

Une gêne que peut rencontrer l'utilisateur de ces logiciels pour des modèles sophistiqués est qu'il est très difficile de s'assurer que le modèle qu'on a spécifié est bien le modèle que l'on voulait utiliser. Ce n'est pas un langage de programmation, on ne peut donc pas effectuer des sorties intermédiaires ou capitaliser des instructions au travers de subroutines. C'est l'origine du projet ReBaStaBa (Réseaux Bayésiens traités par Statistique Bayésienne) [b], paquet R en cours d'élaboration, où les réseaux bayésiens sont définis par des objets R associés à des fonctions pour en visualiser les propriétés (impression, représentation graphique, exploration des parentés,...), les manipuler (construction de sous-réseaux bayésiens, modification de la distribution de probabilité de quelques noeuds,...), les exporter dans des formats utilisés par d'autres paquets centrés sur les réseaux bayésiens comme Rjags, Deal, Grappa.

Bibliographie

[a] Albert, I., Grenier, E., Denis, J.-B. and Rousseau, J., (2008) Quantitative risk assessment from farm to fork and beyond: a global Bayesian approach concerning food-borne diseases., *Risk Analysis*, 28, 557–571.

[b] Denis, J.-B., (2008) Rebastaba project, `http://w3.jouy.inra.fr/unites/miaj/public/matrisq/jbdenis/outils/welcome.html`.

[c] Plummer, M., (2008) Jags, Just Another Gibbs Sampler, `http://www-fis.iarc.fr/~martyn/software/jags/`.

[d] R Development Core Team, (2007) R: A Language and Environment for Statistical Computing, `http://www.R-project.org`.

[e] Thomas, A., (2008) OpenBugs, Bayesian inference Using Gibbs Sampling, `http://mathstat.helsinki.fi/openbugs/`.