Elicitation de loi a priori par combinaison d'experts

S. Donnet [3].
en collaboration avec I. Albert[1], S.L. Choy[2], C. Guihenneuc[4], K. Mengersen[2], J. Rousseau[3]

[1]Unité meta@risk, INRA, Paris

[2]Queensland University of technology, Brisbane, Australia

[3]Université Paris-Dauphine, Cérémade

[4]Université Paris-Descartes, MAP5 et Inserm U754

Journées MAS, Rennes, 27-29 août 2008

Introduction

Définition

"Eliciter" est l'action d'aider un expert à formaliser ses connaissances pour permettre de les sauvegarder et/ou les partager.

- **Contexte** : modélisation bayésienne
- **But** : Construction de loi a priori à partir de dires d'experts
- Problème largement étudié dans le passé (d'un point de vue psychologique et statistique)
- **Ici** : proposer un modèle statistique cohérent combinant les avis de plusieurs experts et construire une loi a priori.

Contexte Bayésien et loi a priori

- Données Y_d avec modèle paramétrique $Y_d \sim \mathcal{M}_\theta$ où $\theta \in \mathbb{R}^d$ sont les paramètres.
- Inférence bayésienne repose sur $\pi(\theta|Y_d)$ sachant que

$$\pi(\theta|Y_d) \propto \pi(Y_d|\theta)\pi(\theta)$$

avec $\pi(Y_d|\theta)$ vraisemblance des données et $\pi(\theta)$ loi a priori sur les paramètres

- Choix de la loi a priori
  - **Pas d'information sur θ** (sauf sur le support) : Loi a priori "plate" (Jeffreys par ex)
  - **Etudes antérieures** : Les lois a posteriori des études antérieures servent de loi a priori
  - **Prise en compte de l'avis d'experts** pour constuire une loi a priori $\pi(\theta)$

Exemple en risque alimentaire

- Y_d = nombre de souris mortes sur n souris ayant reçu une même dose d d'agent pathogène (listeria) donc

$$Y_d \sim \mathcal{B}in(n, p(d, \alpha))$$

où $p(d, \alpha)$ la probabilité de mourir
- Modèle exponentiel : $p(d, \alpha) = 1 - e^{-\alpha d}$
- On cherche à estimer $\theta = \alpha$
-
  - Très peu de données sont disponibles dans ce cas
  - donc utilisation les avis d'experts (biologistes par ex...) pour pallier à cette lacune
  - i.e. construction d'une loi a priori informative $\pi(\theta)$ à partir de dires d'experts

Difficultés et restriction

Données élicitées

Notons $D = \{D_1, \ldots, D_N\}$ les données fournies par les N experts. Ces données sont appelées données élicitées.

- **Difficultés**
  - Difficile pour l'esprit humain de quantifier un savoir qualitatif.
  - Souvent les experts ne fournissent qu'une petite quantité de données

  $\Rightarrow$ Raisonnable de supposer que $\pi(\theta)$ appartient à une famille paramétrique

$$\pi(\theta) \in \{\pi(\theta|\lambda), \lambda \in \Lambda \subset \mathbb{R}^p\}$$

- **But** : sélectionner la valeur de λ la plus appropriée en fonction du savoir des experts
- **Exemple** :
  - $Y_d \sim \mathcal{B}in(n, 1 - e^{-\alpha d}) \Rightarrow \theta = \alpha$
  - $\log(\alpha) \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \lambda = (\mu, \sigma^2)$

Collection de données élicitées

- Processus
  1. **Préparation** : Sélection et motivation des experts
  2. **Définition** des quantités à éliciter
  3. **Interrogation** des experts
  4. **Traitement statistique** : estimation de λ à partir des données élicitées
  5. **Validation** auprès des experts
- Nature des données élicitées
  - But : encoder le savoir des experts en distributions de probabilité.
  - Aperçu plus cohérent du savoir des experts obtenu en utilisant 2 types de questions : par ex. probabilités et quantiles.

Exemple en risque alimentaire : élicitation de probabilités

Questions structurelles

Dans combien d'expériences parmi 100 obtiendrez-vous moins de q% de souris mortes ? avec $q = 20\%$ ou 60%.

- **Réponses théoriques** Pour chaque expert i :
  $\boxed{P_{i,q} = \mathbb{P}(p(d,\alpha) \leq q | \mu_i, \sigma_i^2)}$
  avec $\log \alpha \sim \mathcal{N}(\mu_i, \sigma_i^2)$
- **Traitement statistique**

$$\left\{ \begin{array}{rcl} P_{i,q_1} & = & \mathbb{P}(p(d,\alpha) \leq q_1 | \mu_i, \sigma_i^2) \\ P_{i,q_2} & = & \mathbb{P}(p(d,\alpha) \leq q_2 | \mu_i, \sigma_i^2) \end{array} \right.$$

On en déduit $\lambda_i = (\mu_i, \sigma_i)$ avec

$$\left\{ \begin{array}{rcl} \sigma_i & = & \frac{-\log(1-q_1)+\log(1-q_2)}{d(\Phi^{-1}(P_{i,q_1})-\Phi^{-1}(P_{i,q_2}))} \\ \mu_i & = & \sigma_i \Phi^{-1}(P_{i,p_1}) - \frac{\log(1-q_1)}{d} \end{array} \right.$$

$\Rightarrow$ Quantités non observables donc difficile pour les experts.

Exemple en risque alimentaire : élicitation de probabilités

ou bien... Questions prédictives

Une expérience consiste en l'injection de la dose d à 10 souris. Dans combien d'expériences parmi 100 obtiendrez-vous moins de q souris mortes parmi les 10 ? avec $q = 2$ ou 6.

- **Réponses théoriques** : Pour chaque expert i...
  $P_{i,q} = \mathbb{P}(Y_d \leq q|\mu_i, \sigma_i^2)$ où $Y_d \sim \mathcal{B}in(10, p(d, \alpha))$ avec $\log \alpha \sim \mathcal{N}(\mu_i, \sigma_i^2)$
- Quantités observables mais implique des difficultés calculatoires au moment du traitement

Exemple en risque alimentaire : élicitation de quantiles

Question prédictive

Une expérience consiste en l'injection d'une dose d à 10 souris. Quel serait le nombre q de souris mortes parmi les 10 tel que $t\%$ des expériences aboutissent à un nombre de souris mortes inférieur ou égal à q ? avec $t = 0.1, 0.5, 0.9$

- **Réponses théoriques** $Q_{i,t}$ tel que $\mathbb{P}(Y_d \leq Q_{i,t}|\mu_i, \sigma_i^2) = t$ où $Y_d \sim \mathcal{B}in(10, p(d, \alpha))$ avec $\log \alpha \sim \mathcal{N}(\mu_i, \sigma_i^2)$

ou bien... Question structurelle

Quelle serait la proportion q de souris mortes telle que $t\%$ des expériences aboutissent à une proportion de souris mortes inférieure ou égale à q ? avec $t = 0.1, 0.5, 0.9$

- **Réponses théoriques** $Q_{i,t}$ tel que $\mathbb{P}(p(d, \alpha) \leq Q_{i,t}|\mu_i, \sigma_i^2) = t$ où $\log \alpha \sim \mathcal{N}(\mu_i, \sigma_i^2)$

Erreur et vraisemblance des données élicitées

A chaque quantité D_i, les experts associent un degré de confiance c_i compris entre 1 et 10

Donc on peut construire un modèle d'erreur du type Probit :

$$\begin{cases} \Phi^{-1}(D_i) & = & \Phi^{-1}(D(\lambda_i)) + \varepsilon_i \\ \varepsilon_i & \sim & \mathcal{N}(0, v_i) \end{cases}$$

- $D(\lambda_i)$ est le **quantile / probabilité théorique**. Dans notre exemple : $Q_t(\lambda_i)$ est tel que $\mathbb{P}(p(d,\alpha) \leq Q_t(\lambda_i) | \mu_i, \sigma_i^2) = t$ d'où
$$Q_t(\lambda_i) = 1 - e^{-d \exp(\mu_i + \sigma_i \Phi(t))}$$
- Φ est la f.d.r. d'une $\mathcal{N}(0,1)$ et v_i est relié à la confiance et la cohérence q^* (fixé par le statisticien) par :
$$\mathbb{P}(-q^* \leq \varepsilon_i \leq q^*) = c_i \Leftrightarrow \sqrt{v_i} = \frac{q^*}{\Phi^{-1}\left(\frac{c_i+1}{2}\right)}$$

De là on déduit la vraisemblance des données élicitées $D_1, \ldots, D_N$ en fonction des paramètres λ_i

Combinaison des experts

Soient N experts interrogés. De chaque expert i, on déduit (après traitement statistique) un $\widehat{\lambda}_i$ donc une loi a priori

$$\forall i, \pi(\theta|\widehat{\lambda}_i) \rightarrow \pi(\theta)?$$

1. Approche par mélange $\boxed{\pi(\theta) = \sum_{i=1}^N \omega_i \pi(\theta|\widehat{\lambda}_i)}$
   Permet de modéliser les différences entre experts
2. Approche par moyenne
   $\boxed{\widehat{\lambda} = \sum_{i=1}^N \omega_i \widehat{\lambda}_i \quad \Rightarrow \quad \pi(\theta) = \pi(\theta|\widehat{\lambda})}$
   Permet de trouver un consensus entre experts

Difficultés

Comment choisir les poids ω_i ? Comment prendre en compte les interactions entre experts ?

A propos des interactions entre experts

Les experts peuvent avoir des sources d'information similaires (mêmes expérimentations, mêmes instituts) $\Rightarrow$ devrait fournir le même genre de valeur.

- Si deux experts sont du même institut, les valeurs qu'ils fournissent ont-elles deux fois plus de chances d'être correctes ?
- Sont-elles aussi informatives que celles d'experts indépendants ?

Modèle hiérarchique

Proposer un modèle permettant de prendre en compte leurs savoir commun $\Rightarrow$ Considérer des groupes d'experts avec une structure hiérarchique

Formulation du modèle hiérarchique

Soient K groupes d'experts (groupe = institut, laboratoire, équipe...).
ij est le i-ème expert du groupe j.

Modèle hiérarchique

$$
\begin{array}{ccl}
\lambda_{ij} & \overset{\text{i.i.d.}}{\sim} & g(\cdot|\lambda_j, b_j) \quad \forall j = 1, \ldots, K, \\
\lambda_j & \overset{\text{i.i.d.}}{\sim} & g(\cdot|\lambda, b), \\
\lambda & \sim & \pi_0
\end{array}
$$

- Experts regroupés en classes d'homogénéité avec même distribution. b_j=hyper-paramètre et λ_j=paramètre de centrage
- Savoir des groupes lié par une distribution commune (b= hyper-paramètre et λ= paramètre de centrage)
- π_0 loi peu informative représentant l'incertitude globale. λ peut être vu comme le paramètre de consensus entre les experts

Traitement statistique

"Loi a priori proposée $\pi(\theta)$"= "loi a posteriori $\pi(\theta|D)$"

- **De la forme** :

$$\pi(\theta|D) \propto \int \pi(\theta|\lambda) \prod_{ijt} \underbrace{L(D_{ij}|\lambda_{ij})}_{\text{vrais. des don. élicit.}} \prod_{ij} g(\lambda_{ij}|\lambda_j, b_j) \dots$$

$$\dots \prod_j g(\lambda_j|\lambda, b)\pi_0(\lambda) d\lambda_{ij} d\lambda_j d\lambda$$

- **Estimation** (en utilisant les 2 jeux de données élicitées séparemment) :
  - Hyper-paramètres b_j et b : obtenus par plug-in en utilisant le type 1 de données élicitées (les probabilités par ex.)
  - Loi a priori élicitée $\pi(\theta|D)$ approchée par un algorithme MCMC en utilisant le type 2 de données élicitées (les quantiles par ex.)

Objectifs de l'étude sur données simulées

- Semblable à une analyse de sensibilité : permet de comprendre les hypothèses faîtes dans le modèle.
- Variation de quantités telles que :
  - le nombre d'experts
  - la composition des groupes d'experts
  - la variabilité entre les groupes et au sein des groupes
  - ...

  étude de l'impact de ces variations
- Comparaison des différentes méthodes de combinaison d'experts (moyenne, mélange et modèle hiérarchique)

Les données simulées : paramètres

On se place dans le modèle exponentiel en risque alimentaire.

- $\log \alpha \sim \mathcal{N}(\mu, \sigma^2)$
- Modèle hiérarchique sur les moyennes

$$\mu_{ij} = \mu_j + \Delta_{ij} \text{ avec } \Delta_{ij} \sim \mathcal{N}(0, \tau_j)$$
$$\mu_j = \mu + \Delta_j \text{ avec } \Delta_j \sim \mathcal{N}(0, \tau)$$

- Modèle hiérarchique sur les variances

$$\sigma_{ij}^2 = \sigma_j^2 \rho_{ij} \text{ avec } \rho_{ij} \sim \Gamma(\xi_j, \xi_j)$$
$$\sigma_j^2 = \sigma^2 \rho_j \text{ avec } \rho_j \sim \Gamma(\xi, \xi)$$

- Nous simulons 2 groupes d'experts d'effectifs N_1 et N_2.
- Simulation de 2 probabilités (à 20% et 60%) et 5 quantiles $(0.1, 0.9, 0.25, 0.75, 0.5)$ avec **erreurs**.

Données simulées : phase d'estimation

1. Estimation des variabilités intra-groupes et inter-groupes (hyperparamètres)

$$\tau_j, \xi_j, \tau, \xi$$

   par une phase de **plugin** à partir des probabilités élicitées p_{ijq}

2. Approximation de la loi a priori obtenue par elicitation

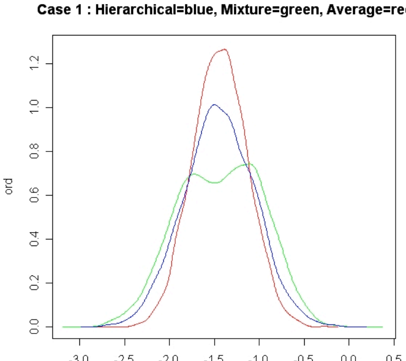
$$\pi(\alpha|\ D)$$

   en utilisant un algorithme MCMC reposant sur les quantiles q_{ijt}

Données simulées (1)

Situation 1 :

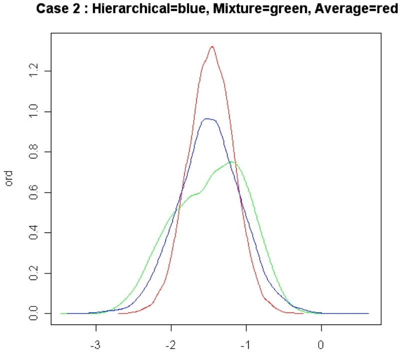
- Même nombre d'experts dans chaque groupe $N_1 = N_2 = 5$
- $\tau_1 = \tau_2 = 0.01 \quad , \quad \frac{1}{\xi_1} = \frac{1}{\xi_2} = 0.01$ (groupes homogènes)
- $\mu_1 = -1.8, \ \mu_2 = -1.1$
- Comparaison des 3 méthodes : tracés des densités des lois a priori $\pi(\alpha|D)$ obtenues par les 3 méthodes



Données simulées (2)

Situation 2 :

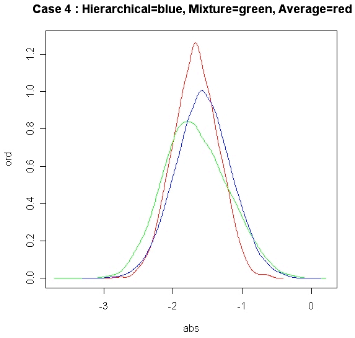
- Même nombre d'experts dans chaque groupe $N_1 = N_2 = 5$
- Groupes inhomogènes : $\tau_1 = 0.1 \quad , \quad \tau_2 = 0.01$
  $\frac{1}{\xi_1} = \frac{1}{\xi_2} = 0.01$
- $\mu_1 = -1.8$, $\mu_2 = -1.1$
- Comparaison des 3 méthodes :



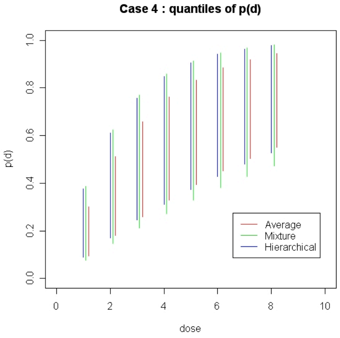
Données simulées (3)

Situation 3 :

- Groupes déséquilibrés en effectifs $N_1 = 10,\ N_2 = 2$
- Groupes homogènes : $\tau_1 = \tau_2 = 0.01 \quad , \quad \frac{1}{\xi_1} = \frac{1}{\xi_2} = 0.01$
- $\mu_1 = -1.8,\ \mu_2 = -1.1$
- Comparaison des 3 méthodes :



Comparaison des densités



Comparaison des intervalles de confiance de $p(d, \alpha)$ en fonction de la dose

Conclusion et travaux en cours

- Méthode générique avec cadre statistique cohérent.
- Applications sur données réelles et simulées :
  - Deux modèles de risque alimentaire
  - Un modèle sur la durée des thèses en mathématiques appliquées dans une université australienne.
- En cours :
  - **Sur données simulées :**
    - Robustesse de la méthode. Ex : erreur sur le nbre de groupes...
    - Mise en place d'un critère de comparaison (autre que graphique)
  - **Sur données réelles**
    - Validation avec les experts i.e vérification par les experts de l'adéquation des distributions avec son ressenti
    - Comparaison sur données réelles entre une loi apriori plate et notre loi a priori.
    - ...