Examen (2h)

28 mars 2017

Toutes les réponses doivent être justifiées soigneusement, en particulier pour les questions auxquelles on peut répondre par oui ou par non. Sans justification, une réponse juste ne rapportera pas de point. Les deux exercices sont indépendants et comptent environ pour la moitié de la note chacun.

Exercice 1

Nous considérons le modèle de régression linéaire Gaussien suivant :

$$Y_i|x_i,\beta,\sigma^2 \sim \mathcal{N}(x_i^\intercal\beta,\sigma^2), \forall i \in \{1,\ldots,n\},$$

où $\text{support}(Y_i) = \mathbb{R}$, $x_i \in \mathbb{R}^p$, $\beta \in \mathbb{R}^p$ et $\sigma \in \mathbb{R}^+$. De plus, le vecteur des poids de la régression est caractérisé par une loi *a priori* Gaussienne :

$$p(\beta|\alpha) = \mathcal{N}(0_p, \frac{1}{\alpha}),$$

où $\alpha \in \mathbb{R}^{+*}$. Nous disposons de la réalisation d'un n-échantillon $\{(x_1,y_1),\ldots,(x_n,y_n)\}$ et l'objectif premier est d'estimer β à partir des données. La variance du bruit σ^2 est ici supposée connue. En revanche, le paramètre α (appelé hyperparamètre dans la littérature) contrôlant l'inverse variance de la loi *a priori* doit également être estimé. Pour simplifier les expressions, nous noterons $X = (x_{ij})_{ij}$ la matrice dont la ligne i correspond à $x_i^\intercal$. De la même manière, nous noterons Y le vecteur construit à partir des y_i.

Question 1 Montrez par identification que la loi *a posteriori* de β sachant l'échantillon, σ^2, α, est donnée par

$$p(\beta|(x_1,y_1),\ldots,(x_n,y_n),\sigma^2,\alpha) = \mathcal{N}(\beta; m_n; S_n), \tag{1}$$

où $S_n^{-1} = X^\intercal X/\sigma^2 + \alpha I_p$ et $m_n = S_n X^\intercal Y/\sigma^2$.

Question 2 Donnez l'expression exacte de la log vraisemblance marginale

$$\log p(y_1,\ldots,y_n|x_1,\ldots,x_n,\sigma^2,\alpha),$$

β est intégré.

Nous cherchons maintenant à dériver un algorithme EM permettant de maximiser la log vraisemblance marginale.

Question 3 Montrez que la log vraisemblance des données complétées s'écrit :

$$\log p(y_1,\ldots,y_n,\beta|x_1,\ldots,x_n,\sigma^2,\alpha) = -\frac{1}{2\sigma^2}||Y - X\beta||^2 - \frac{\alpha}{2}\beta^\intercal\beta + \frac{p}{2}\log(\alpha) + \text{const}, \tag{2}$$

où const ne dépend ni de β ni de α.

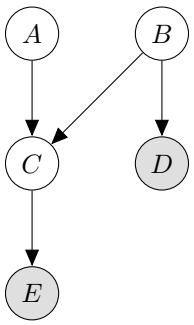
Question 4 Déterminez l'espérance de la log vraisemblance des données complétées (2) à partir de la loi *a posteriori* (1).

Question 5 Etape M : donnez l'estimateur de α maximisant l'espérance de la log vraisemblance des données complétées obtenue à la question précédente.

Question 6 Donnez l'algorithme EM complet permettant de maximiser la log vraisemblance marginale des données.

Exercice 2

On étudie le modèle graphique orienté suivant :



Question 1 Écrire la condition de factorisation que doit vérifier une loi sur les variables aléatoires A, B, C, D et E (supposées discrètes) pour être compatible avec le modèle graphique.

Question 2 Indiquer pour chaque indépendance conditionnelle suivante si elle est bien vérifiée dans le modèle en justifiant brièvement chaque réponse.

$$A \perp\!\!\!\perp B \tag{3}$$
$$A \perp\!\!\!\perp B \mid E \tag{4}$$
$$B \perp\!\!\!\perp E \mid C \tag{5}$$
$$C \perp\!\!\!\perp D \mid A \tag{6}$$

Question 3 Pour une loi jointe compatible avec le modèle graphique, donner $p(e|a,b)$ sous forme d'une expression faisant intervenir la somme sur les valeurs c de C et des lois conditionnelles naturelles dans le modèle graphique étudié.

On suppose à partir de maintenant que les variables aléatoires étudiées sont toutes binaires (à valeurs dans $\{0,1\}$).

Question 4 Indiquer l'ensemble des paramètres nécessaires pour spécifier complètement une loi jointe $p(A,B,C,D,E)$ compatible avec le modèle graphique, en précisant le rôle de chaque paramètre et les différentes lois (conditionnelles) utilisées. On notera θ le vecteur constitué de l'ensemble des paramètres considérés.

Question 5 Écrire la vraisemblance complétée $p(a,b,c,d,e|\theta)$ sous le modèle paramétrique décrit dans la question précédente.

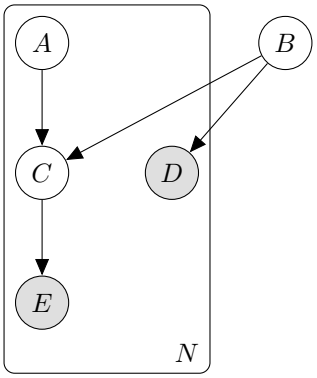
On suppose maintenant qu'on dispose de N observations $(d_i,e_i)_{1\leq i\leq N}$ indépendantes et identiquement distribuées selon la loi marginale associée au modèle graphique. On cherche à estimer le paramètre θ par maximum de vraisemblance à partir des observations.

Question 6 Écrire la vraisemblance des paramètres pour une observation, puis à partir de celle-ci, la vraisemblance pour l'ensemble des observations.

Question 7 Combien de configurations des variables cachées faut-il évaluer pour calculer la vraisemblance ?

Question 8 Pourrait-on utiliser l'algorithme EM pour estimer θ par maximum de vraisemblance dans cette situation ? Si oui, indiquer quelles lois à posteriori doivent être calculées (sans faire les calculs).

Questions bonus On considère à partir de maintenant le modèle suivant :



Question 9 Développer la notation par plaque pour $N = 2$.

Question 10 Donner la forme générale des lois compatibles avec le modèle graphique (pour N quelconque).

Question 11 Par quoi faut-il conditionner les C_i pour les rendre indépendants (on donnera le conditionnement faisant intervenir le moins de variables possible) ?