Modèles à variables latentes pour l'écologie

Examen du 30 mars 2026

Durée : 2 heures

Les notes de cours sur papier et une calculatrice sont autorisées,
à l'exclusion de tout autre appareil électronique (tablette & téléphone compris).

1 Modèle de mélange Béta-Uniforme

Modèle. On considère le modèle de mélange de deux lois Béta suivant :

$$\begin{aligned} &(Z_i)_{1\leq i\leq n} \sim \mathcal{B}(\pi), \\ (Y_i)_{1\leq i\leq n} \text{ indépendants} \mid (Z_i) : \quad &(Y_i \mid Z_i = 0) \sim \mathrm{B}(1,1), \\ &(Y_i \mid Z_i = 1) \sim \mathrm{B}(1,\beta), \end{aligned} \tag{1}$$

où $\mathcal{B}(\cdot)$ désigne la loi de Bernoulli, $\mathrm{B}(\cdot,\cdot)$ désigne la loi Béta et dans lequel la variable Z n'est pas observée. On cherche ici à estimer par maximum de vraisemblance le paramètre $\theta = (\pi, \beta)$ au moyen d'un algorithme EM.

Rappel. On rappelle que la densité de la loi $\mathrm{B}(\alpha,\beta)$ est définie pour $x \in [0,1]$ et vaut

$$\mathrm{B}(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

La loi $\mathrm{B}(1,1)$ est donc la loi uniforme sur $\mathcal{U}[0,1]$.
On rappelle de plus que, pour $x > 0$, la fonction Γ vérifie $\Gamma(x+1) = x\Gamma(x)$.

Questions.

1. Écrire la log-vraisemblance complète $\log p_\theta(Y,Z)$ du modèle (1).

Solution. On a

$$\begin{aligned} \log p_\theta(Y,Z) &= \log p_\theta(Z) + \log p_\theta(Y \mid Z) \\ &= \sum_{i=1}^n \left[Z_i \log \pi + (1-Z_i)\log(1-\pi)\right] + \sum_{i=1}^n Z_i \log \mathrm{B}(Y_i;1,\beta), \end{aligned}$$

puisque $\mathrm{B}(Y_i;1,1) = 1$ donc $\log \mathrm{B}(Y_i;1,1) = 0$, donc

$$\begin{aligned} \log p_\theta(Y,Z) &= \sum_{i=1}^n \left[Z_i \log \pi + (1-Z_i)\log(1-\pi)\right] \\ &\quad + \sum_{i=1}^n Z_i \left[(\beta-1)\log(1-Y_i) + \log\Gamma(1+\beta) - \log\Gamma(\beta) - \log\Gamma(1)\right], \\ &= \sum_{i=1}^n \left[Z_i \log \pi + (1-Z_i)\log(1-\pi)\right] + \sum_{i=1}^n Z_i \left[(\beta-1)\log(1-Y_i) + \log\beta\right], \end{aligned}$$

puisque $\Gamma(1) = 1$ et $\Gamma(1+\beta) = \beta\Gamma(\beta)$.

2. Écrire explicitement l'étape E de l'algorithme EM associé.

Solution. Pour la valeur courante du paramètre $\theta^{(h)} = (\pi^{(h)}, \beta^{(h)})$, il s'agit d'évaluer l'espérance conditionnelle $Q(\theta \mid \theta^{(h)}) = \mathbb{E}_{\theta^{(h)}}[\log p_\theta(Y, Z) \mid Y]$, c'est-à-dire

$$Q(\theta \mid \theta^{(h)}) = \sum_{i=1}^{n} \left[\tau_i^{(h)} \log \pi + (1 - \tau_i^{(h)}) \log(1 - \pi) + \tau_i^{(h)} \left((\beta - 1) \log(1 - Y_i) + \log \beta \right) \right]$$

où

$$\begin{aligned} \tau_i^{(h)} &= \mathbb{E}_{\theta^{(h)}}[Z_i \mid Y] = \mathbb{P}_{\theta^{(h)}}\{Z_i = 1 \mid Y\} = \mathbb{P}_{\theta^{(h)}}\{Z_i = 1 \mid Y_i\} \\ &= \frac{\pi^{(h)} \mathrm{B}(Y_i; 1, \beta^{(h)})}{(1 - \pi^{(h)}) + \pi^{(h)} \mathrm{B}(Y_i; 1, \beta^{(h)})}. \end{aligned}$$

3. Écrire explicitement l'étape M de l'algorithme EM associé.

Solution. Il s'agit maintenant de trouver $\theta^{(h+1)} = (\pi^{(h+1)}, \beta^{(h+1)})$ qui maximise $Q(\theta \mid \theta^{(h)})$ par rapport à θ. On calcule pour cela les dérivées

$$\begin{aligned} \partial_\pi Q(\theta \mid \theta^{(h)}) &= \sum_{i=1}^{n} \left[\tau_i^{(h)} / \pi - (1 - \tau_i^{(h)}) / (1 - \pi) \right] = \frac{\sum_{i=1}^{n} \tau_i^{(h)}}{\pi(1 - \pi)} - \frac{n}{1 - \pi}, \\ \partial_\beta Q(\theta \mid \theta^{(h)}) &= \sum_{i=1}^{n} \tau_i^{(h)} \left[\log(1 - Y_i) + 1/\beta \right] \end{aligned}$$

qui s'annulent respectivement pour

$$\pi^{(h+1)} = \frac{1}{n} \sum_{i=1}^{n} \tau_i^{(h)}, \qquad \beta^{(h+1)} = - \left(\sum_{i=1}^{n} \tau_i^{(h)} \right) \Bigg/ \left(\sum_{i=1}^{n} \tau_i^{(h)} \log(1 - Y_i) \right).$$

2 Interaction entre espèces de mouches à la Réunion

Cet exercice porte principalement sur l'interprétation de résultats d'analyses. Les réponses doivent être argumentées, tout en restant synthétiques.

On s'intéresse à la distribution de six espèces de mouches du fruit dans l'île de la Réunion. Il s'agit notamment d'étudier les effets sur les abondances des différentes espèces

— de covariables environnementales,
— de l'espèce de plante sur lesquelles elles sont observées et
— des possibles interactions que ces espèces entretiennent entre elles.

On a relevé pour cela les abondances de $p = 6$ espèces de mouches (*zona, rosa, capi, cucurbitae, ciliatus* et *demmerezi*) sur $n = 4475$ plantes issues de 21 espèces, en relevant à chaque fois 10 covariables environnementales (altitude, niveau de pluie, température, ...), réunies dans un vecteur noté x. Pour $\ell = 1, \dots 21$, on définit de plus la variable indicatrice $u_{i\ell}$ qui vaut 1 si le site i correspond à la plante ℓ et 0 sinon.

Modèle Poisson log-normal. Toutes les analyses sont menées dans le cadre du modèle Poisson log-normal (PLN) qu'on rappelle ici. On note $i = 1, \dots, n$ le site de collecte, $j = 1, \dots, p$ l'espèce de mouche et Y_{ij} le nombre d'individus de l'espèce j observés dans le site i. Le modèle PLN suppose alors qu'un vecteur latent gaussien de dimension p est associé à chaque site et que l'abondance de chaque espèce dans chaque site dépend de l'environnement dans le site et de la coordonnée correspondante du vecteur gaussien :

$$\begin{aligned} (Z_i)_{1 \leq i \leq n} \text{ iid} : &\qquad Z_i = [Z_{i1}, \dots, Z_{ip}]^\top \sim \mathcal{N}_p(0, \Sigma), \\ (Y_{ij})_{1 \leq i \leq n, 1 \leq j \leq p} \text{ indépendants } \mid Z : &\qquad Y_{ij} \mid Z_{ij} \sim \mathcal{P}(\exp(\mu_{ij} + Z_{ij})). \end{aligned} \tag{2}$$

	(a)	(b)	(c)	(d)	(e)	(f)
d	1	11	21	31	31	31
D	27	87	147	207	192	197
$ELBO$	−31654	−29196	−24773	−24102	−24143	−24105
BIC	−31767	−29562	−25391	−24972	−24950	−24933

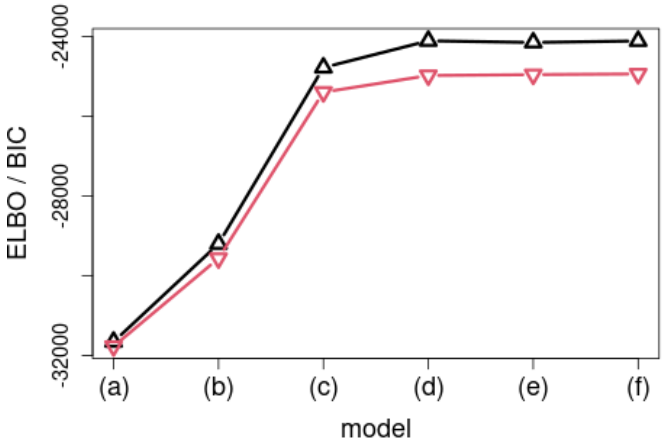

FIGURE 1 – ELBO et critère BIC pour chacun des sept modèles. Haut : valeurs numériques des critères (d désigne le nombre de régresseurs et D le nombre total de paramètres indépendants). Bas : $\triangle$ = ELBO, $\triangledown$ = BIC.

On rappelle que l'inférence de ce modèle peut s'effectuer au moyen d'une approximation (variationnelle) de la loi conditionnelle $p_\theta(Z \mid Y)$ par une loi normale multivariée notée $q_\psi(Z)$ et que l'algorithme VEM associé vise à maximiser la borne inférieure de la vraisemblance (notée ELBO) définie par

$$J(\theta, \psi) = \log p_\theta(Y) - KL[q_\psi(z)||p_\theta(Z \mid y)].$$

Modèles considérés. Pour répondre aux questions énoncées plus haut, on considère les sept modèles suivants (en l'absence de précision, aucune contrainte n'est appliquée sur la matrice Σ, c'est-à-dire que la matrice est supposée pleine) :

(a)	$\mu_{ij} = 1$	aucune effet, constante seule ($d = 1$),
(b)	$\mu_{ij} = \beta_0 + x_i^\top \beta_j$	covariables environnementales plus constante ($d = 11$),
(c)	$\mu_{ij} = \sum_{\ell=1}^{21} u_{i\ell}\alpha_\ell^j$	indicatrice de plante ($d = 21$),
(d)	$\mu_{ij} = \sum_{\ell=1}^{21} u_{i\ell}\alpha_\ell^j + x_i^\top \beta_j$	covariables environnementales et indicatrice de plante ($d = 31$),
(e)	$\mu_{ij} = \sum_{\ell=1}^{21} u_{i\ell}\alpha_\ell^j + x_i^\top \beta_j$	modèle (d) avec Σ diagonale (aucune interaction biotique),
(f)	$\mu_{ij} = \sum_{\ell=1}^{21} u_{i\ell}\alpha_\ell^j + x_i^\top \beta_j$	modèle (d) avec $\Omega = \Sigma^{-1}$ creuse (quelques interactions biotiques),

où β_j est le vecteur des coefficients de régression des covariables environnementales x_i pour l'espèce j, α_ℓ^j est l'effet de la plante ℓ sur l'abondance de l'espèce j et en notant d le nombre total de covariables incluses dans le modèle.

Résultats. Les figures 1, 2 et 3 présentent les résultats obtenus avec ces différents modèles. La figure 1 compare les ajustements des différents modèles, la figure 2 présente les estimations des effets des différentes plantes sur l'abondance des différentes espèces de mouche dans le modèle (d) et la figure 3 présente le réseau d'interactions biotiques estimé dans le modèle (f).

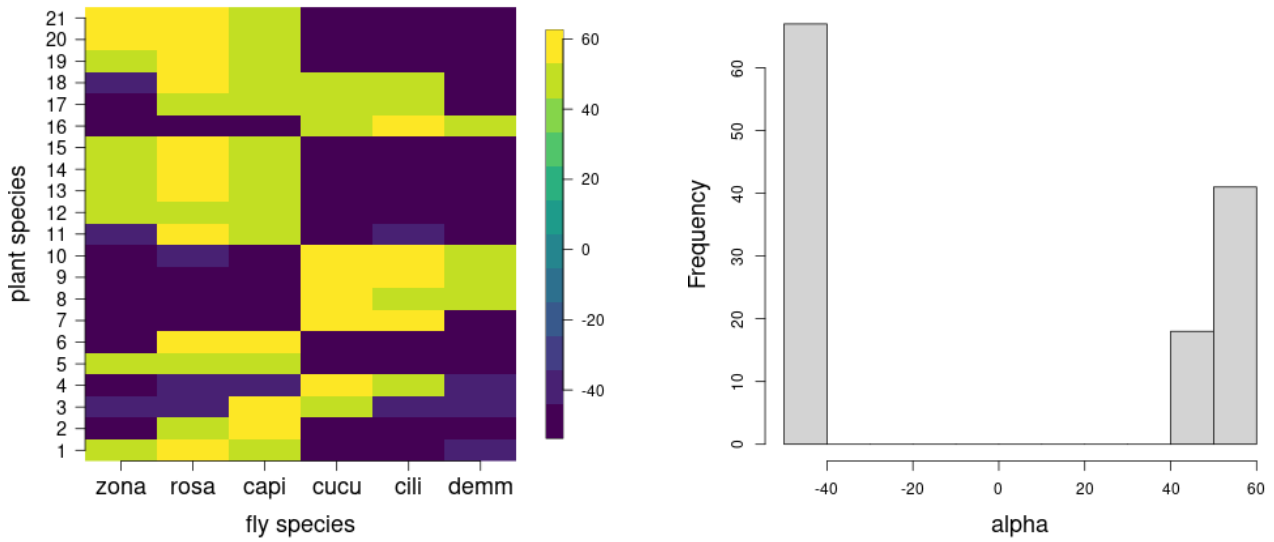

FIGURE 2 – Effets α_ℓ^j des 21 plantes sur chacune des 6 espèces de mouches selon le modèle (d). Gauche : matrice des coefficients estimés $\widehat{\alpha} = [\widehat{\alpha}_\ell^j]_{1\leq\ell\leq 21, 1\leq j\leq 6}$ en code couleur (plantes en lignes, mouches en colonnes). Droite : distribution de l'ensemble des $21 \times 6 = 126$ coefficients estimés $\{\widehat{\alpha}_\ell^j\}_{1\leq\ell\leq 21, 1\leq j\leq 6}$.

	zona	rosa	capi	cucu.	cili.	demm.
zona	0.17	0.03	0.01	0	0	0
rosa	0.03	0.63	0	0	0	0
capi	0.01	0	0.29	0	0	0
cucu.	0	0	0	0.19	0.02	0.01
cili.	0	0	0	0.02	0.22	0.01
demm.	0	0	0	0.01	0.01	0.18

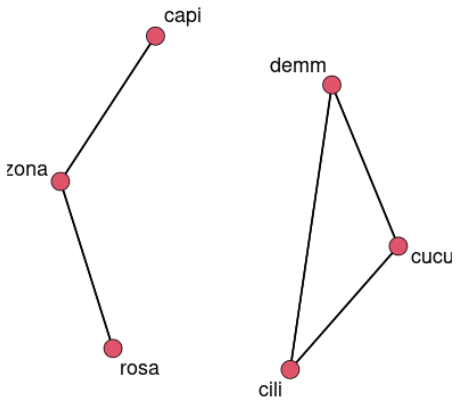

FIGURE 3 – Réseau d'interactions biotiques estimé selon le modèle (f). Gauche : matrice de précision latente estimée $\widehat{\Omega} = \widehat{\Sigma^{-1}}$. Droite : réseau estimé (une arête est présente si le terme correspondant dans $\widehat{\Omega}$ est non nul).

Questions.

1. Rappeler la formule du critère BIC (fondé sur l'ELBO) pour le modèle PLN (2) pour n sites, p espèces et d covariables, sans contrainte sur la matrice Σ.

Solution. Par définition, le critère BIC vaut

$$BIC = ELBO - \frac{D}{2}\log(n)$$

où D est le nombre de paramètres indépendants. Dans le cas présent, le modèle (2) implique $p \times d$ coefficients de régressions et $p(p+1)/2$ paramètres de variance-covariances, soit au total

$$D = p[d + (p+1)/2] \text{ paramètres indépendants.}$$

2. Au vu de la figure 1 : qui de l'environnement et de la plante, a l'effet le plus important ? Parmi les modèles sans contrainte sur Σ (soit (a), (b), (c) et (d)), lequel retiendriez-vous finalement ?

Solution. Le critère BIC augmente fortement quand on introduit chacun des deux effets (environnement et plante), mais l'augmentation est beaucoup plus forte pour la plante (BIC(c) - BIC(a) = +5825) que pour l'environnement BIC(b) - BIC(a) = +1654).
Au total, le modèle (d) incluant les deux effets obtient le meilleur critère BIC (BIC(d) = -24972) et semble donc néanmoins le plus pertinent.

3. Commenter les valeurs des effets estimés des plantes surt les mouches présentés dans la figure 2. Que pouvez-vous dire des préférences des trois premières espèces de mouches (zona, rosa et capi) en matière de plante par rapport à celles des trois suivantes (cucurbitae, ciliatus et demmerezi) ?

Solution. Les valeurs des coefficients sont très contrastées : soit très négatives, soit très positives : les préférences des différents espèces de mouches sont donc très marquées..
Les signes des coefficients sont presque systématiquement opposés entre les espèces du premier groupe et celles du second : les préférences de ces deux groupes d'espèces sont donc opposées (ou complémentaires) et on peut supposer que les espèces issues de ces deux groupes sont rarement observées sur la même plante.

4. En vous fondant seulement sur la figure 1, que pouvez-vous dire de l'intensité des interactions biotiques entre les six espèces de mouches considérées ?

Solution. Le critère BIC du modèle (e), qui ne prévoit aucune interaction, est supérieure à celui du modèle (d), qui les autorise toutes. Le modèle qui suppose que les espèces s'ignorent les unes les autres l'emporte donc.
Le modèle (g) qui suppose l'existence d'un nombre limité d'interactions est néanmoins préférable.

5. Commenter la valeur et le signe des éléments de $\widehat{\Omega}$ présentés dans la figure 3 : à quel type d'interaction peuvent-ils être associés ? Que pouvez-vous dire de la structure du réseau inféré au vu des résultats précédents.

Solution. On sait que Ω_{jk} est signe opposé à la corrélation partielle entre les espèces j et k. Toutes les corrélation partielles étant négatives, on peut penser à des relations de compétition. Ces corrélation sont néanmoins très faibles, donc la compétition également.
Le réseau inféré n'est pas connexe et sépare les même deux groupes de mouches que les effets des différentes plantes : zona, rosa et capi, d'une part, et cucurbitae, ciliatus et demmerezi, d'autre part. On peut donc conclure que ces deux groupes d'espèces ont des habitats presque complètement distincts et qu'ils n'entretiennent donc aucune interaction.

3 Approximation variationnelle de type 'Espérance-Propagation'

Problème. Les algorithmes de type 'Espérance-Propagation' se fondent sur une approximation variationnelle d'une loi cible p par une loi q choisie dans une famille de lois $\mathcal{Q}$ par minimisation de la divergence de Küllback-Leibler :

$$KL(p||q) = \mathbb{E}_p\left[\log\left(\frac{p(X)}{q(X)}\right)\right],$$

où $\mathbb{E}_p$ désigne l'espérance selon la loi p.

On s'intéresse ici à la forme de la loi optimale

$$\widetilde{q} = \arg\min_{q\in\mathcal{Q}} KL(p||q)$$

pour certaines familles de lois $\mathcal{Q}$ particulières.

Rappel. On rappelle que pour une matrice A inversible, et en notant $|A|$ le déterminant de A, on a

$$\partial_A\left(u^\top A u\right) = uu^\top, \qquad \partial_A\left(\mathrm{tr}(AB)\right) = B^\top, \qquad \partial_A\left(\log|A|\right) = \left(A^{-1}\right)^\top.$$

Approximation gaussienne. On suppose tout d'abord de la loi p porte sur $\mathbb{R}^d$ et que $\mathcal{Q}$ est la famille des lois normales multivariées $\mathcal{N}_d(m,S)$. Il s'agit donc de déterminer couple constitué du vecteur d'espérance $\widetilde{m}$ et de la matrice de variance $\widetilde{S}$ optimal pour une loi p donnée.

1. Calculer la divergence de Küllback-Leibler entre une loi p arbitraire et la loi $q = \mathcal{N}_d(m,S)$, en fonction de m et S, en supposant S de plein rang.

Solution. On part de la définition

$$\begin{aligned} KL(p||q) &= \mathbb{E}_p\left[\log(p(X) + \frac{d}{2}\log(2\pi) + \frac{1}{2}\log(|S|) + \frac{1}{2}(X-m)^\top S^{-1}(X-m)\right] \\ &= -\frac{1}{2}\log(|S^{-1}|) + \frac{1}{2}\mathbb{E}_p\left[(X-m)^\top S^{-1}(X-m)\right] + \mathrm{cst} \end{aligned}$$

ou cst ne dépend ni de m, ni de S. En notant $\mu = \mathbb{E}_p X$ et $\Sigma = \mathbb{V}_p X$ et en utilisant la propriété sur les moments vue en cours $\mathbb{E}\left[(X-a)^\top B(X-a)\right] = (\mu-a)B(\mu-a) + \mathrm{tr}(\Sigma B)$, on obtient

$$KL(p||q) = -\frac{1}{2}\log(|S^{-1}|) + \frac{1}{2}(\mu-m)^\top S^{-1}(\mu-m) + \frac{1}{2}\mathrm{tr}(\Sigma S^{-1}) + \mathrm{cst}.$$

2. Montrer que cette divergence est minimale pour $\widetilde{m} = \mathbb{E}_p(X)$ et $\widetilde{S} = \mathbb{V}_p(X)$.
(*On se contentera d'identifier les points stationnaires de la fonction* $(m,S) \to KL(p||q)$.)

Solution. La dérivée par rapport à m vaut

$$\partial_m KL(p||q) = -S^{-1}(\mu - m),$$

qui s'annule seulement pour $\widetilde{m} = \mu$, puisque S est inversible.
La dérivée par rapport à S^{-1} vaut

$$\partial_S KL(p||q) = -\frac{1}{2}S + \frac{1}{2}(\mu-m)(\mu-m)^\top + \frac{1}{2}\Sigma,$$

qui s'annule pour $\widetilde{S} = (\mu-m)(\mu-m)^\top + \Sigma$, soit $\widetilde{S} = \Sigma$ en injectant la valeur optimale $\widetilde{m}$.

3. Quels avantages et inconvénients la divergence $KL(p||q)$ présente-t-elle par rapport à la divergence $KL(q||p)$?

Solution. La divergence $KL(p||q)$ peut sembler préférable à la divergence $KL(q||p)$ dans la mesure où l'erreur $\log\left(p(X)/q(X)\right)$ est moyennée selon la loi cible p et non selon la loi approchée q.

Les paramètres variationnels $\widetilde{m}$ et $\widetilde{S}$ sont naturels puisqu'ils font coïncider les deux premiers moments de la loi approchée q avec ceux de la loi cible p.
Cependant cette approximation nécessite d'évaluer les deux premier moments (espérance et variance) de X sous la loi p qui est, a priori, difficile à manipuler.

Famille exponentielle. On prend maintenant pour $\mathcal{Q}$ la famille des lois exponentielles de la forme

$$q(x) = \exp\left(\psi^\top t(x) - a(x) - b(\psi)\right),$$

où les fonctions a et b sont données (b étant continue, infiniment dérivable et bijective) et $t(x)$ est un vecteur fonction de x : $t(x) = [t_1(x) \dots t_d(x)]^\top \in \mathbb{R}^d$. Pour une loi p donnée, on veut maintenant déterminer le paramètre ψ associé à la loi $\widetilde{q}$ optimale dans $\mathcal{Q}$.

4. Montrer que $\partial_\psi b(\psi) = \mathbb{E}_q[t(X)]$.

Solution. Vu en cours.

5. Montrer que le paramètre optimal $\widetilde{\psi}$ vérifie $\partial_\psi b(\widetilde{\psi}) = \mathbb{E}_p[t(X)]$.

Solution. On part de la définition

$$\begin{aligned} KL(p||q) &= \mathbb{E}_p\left[\log(p(X) - \psi^\top t(X) + a(X) + b(\psi)\right] \\ &= -\psi^\top \mathbb{E}_p[t(X)] + b(\psi) + \text{cst} \end{aligned}$$

ou cst ne dépend pas de ψ. La dérivée par rapport à ψ vaut donc

$$\partial_\psi KL(p||q) = -\mathbb{E}_p[t(X)] + \partial_\psi b(\psi)$$

et s'annule bien si $\partial_\psi b(\psi) = \mathbb{E}_p[t(X)]$.
On vérifie qu'il s'agit bien d'un minimum en remarquant que

$$\partial^2_{\psi^2} KL(p||q) = \partial^2_{\psi^2} b(\psi) = \mathbb{V}_q[t(X)]$$

(la dernière égalité a été vue en cours) : la hessienne est donc bien définie positive.

6. Commenter ce résultat.

Solution. Comme dans le cas gaussien (qui est un cas particulier de la famille exponentielle), l'approximation revient donc à faire coïncider l'espérance de $t(X)$ sous la loi cible p et sous la loi approchée q. Là encore ce calcul peut s'avérer délicat puisque p est, par nature, difficile à manier.