Sélection de variables en régression

Jean-Michel Marin

Institut de Mathématiques et Modélisation
Université Montpellier 2

Paradigme bayésien paramétrique

Soit $f(\mathbf{y}|\theta)$ la vraisemblance du modèle paramétrique considéré.

$\theta \in \Theta$ est un paramètre inconnu.

θ est considéré comme une quantité aléatoire de densité $\pi(\theta)$,
la loi a priori.

L'inférence bayésienne est basée sur la loi a posteriori :

$$\pi(\theta|\mathbf{y}) \propto f(\mathbf{y}|\theta)\pi(\theta)\,.$$

Choix bayésien de modèles

$$\mathcal{M}_1 : \mathbf{y} \sim f_1(\mathbf{y}|\theta_1), \theta_1 \in \Theta_1, \theta_1 \sim \pi_1(\theta_1)\,.$$
$$\mathcal{M}_2 : \mathbf{y} \sim f_2(\mathbf{y}|\theta_2), \theta_2 \in \Theta_2, \theta_2 \sim \pi_2(\theta_2)\,.$$

Munissons l'espace des modèles d'une loi de probabilité a priori : $\mathbb{P}(\mathcal{M}_1)$ et $\mathbb{P}(\mathcal{M}_2)$.

Un choix de modèle bayésien est basé sur la loi a posteriori des différents modèles :

$$\mathbb{P}(\mathcal{M}_i|\mathbf{y}) \propto \mathbb{P}(\mathcal{M}_i) \int_{\Theta_i} f_i(\mathbf{y}|\theta_i)\pi_i(\theta_i)d\theta_i\,.$$

Typiquement, si $\mathbb{P}(\mathcal{M}_1|\mathbf{y}) > 0.5$, on choisira le modèle 1.

Difficultés associées à cette méthodologie :

- La distribution a posteriori des modèles est très sensible au choix de $\pi_1(\theta_1)$ et $\pi_2(\theta_2)$.

  Si l'on dispose d'informations a priori, il est important que ces lois a priori soient équitables. C'est un problème difficile très peu étudié.

- Il n'est pas possible d'utiliser des lois a priori impropres.

  C'est un problème qui a été beaucoup étudié mais, dans de nombreux cas, les réponses apportées ne sont pas satisfaisantes.

- Pour des modèles complexes, nous ne pouvons pas calculer explicitement $\int_{\Theta_i} f_i(y|\theta_i)\pi_i(\theta_i)d\theta_i$.

- Lorsque le nombre de modèles en compétition est très important, il n'est pas possible de calculer explicitement la loi a posteriori des modèles.

  L'exploration de l'espace des modèles peut alors s'avérer très difficile.

Modèle de régression linéaire gaussien

Nous observons le n-échantillon : $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p$.

Modèle $\gamma = (\gamma_1, \ldots, \gamma_p) \in \Gamma = \{0,1\}^{\otimes p}$:

$$\mathbf{y}|\mathbf{X}, \gamma, \beta^\gamma, \sigma^2 \sim \mathcal{N}_{p_\gamma+1}\left(\mathbf{X}^\gamma \beta^\gamma, \sigma^2 I_n\right) ,$$

- $p_\gamma = \sum_{i=1}^p \gamma_i$,
- $\mathbf{X}^\gamma$ la matrice dont les colonnes sont composées du vecteur $\mathbf{1}_n$ et des variables $\mathbf{x}_i$ dont $\gamma_i = 1$ ($\mathbf{X}^{(1,\ldots,1)} = \mathbf{X}$),
- $\beta^\gamma \in \mathbb{R}^{p_\gamma+1}$ et $\sigma^2 \in \mathbb{R}_+^*$ sont les paramètres inconnus.

Objectif : déterminer le modèle le plus pertinent parmi les 2^p modèles en compétition, inférer sur le paramètre γ.

Lois a priori de Zellner compatibles

Pour γ fixé,

$$\beta^\gamma|\mathbf{X},\gamma,\sigma^2 \sim \mathcal{N}_{p_\gamma+1}(\tilde{\beta}^\gamma, g_\gamma\sigma^2((\mathbf{X}^\gamma)'\mathbf{X}^\gamma)^{-1})\,,$$

$$\pi(\sigma^2|\mathbf{X},\gamma) \propto \sigma^{-2}\,.$$

Le modélisateur choisit l'espérance a priori $\tilde{\beta}^\gamma$ et g_γ : g_γ donne la quantité relative d'information a priori par rapport à celle portée par l'échantillon,

$$\mathbb{E}(\beta^\gamma|\mathbf{X},\gamma,\mathbf{y}) = \frac{g_\gamma\hat{\beta}^\gamma + \tilde{\beta}^\gamma}{g_\gamma+1}\,.$$

Compatibilité de $\pi_1(\theta_1)$ et $\pi_2(\theta_2)$: information de Kullback-Leibler entre $f_1(\mathbf{y}) = \int_{\Theta_1} f_1(\mathbf{y}|\theta_1)\pi_1(\theta_1)d\theta_1$ et $f_2(\mathbf{y}) = \int_{\Theta_2} f_2(\mathbf{y}|\theta_2)\pi_2(\theta_2)d\theta_2$.

Plus cette information est faible plus les lois a priori sont jugées équitables.

$$\beta^1|\mathbf{X}^1, \sigma^2 \sim \mathcal{N}_{k_1}\left(\tilde{\beta}^1, \sigma^2 g_1((\mathbf{X}^1)'\mathbf{X}^1)^{-1}\right)$$

$$\beta^2|\mathbf{X}^2, \sigma^2 \sim \mathcal{N}_{k_2}\left(\tilde{\beta}^2, \sigma^2 g_2((\mathbf{X}^2)'\mathbf{X}^2)^{-1}\right)$$

$\mathcal{M}_2$ est un sous-modèle de $\mathcal{M}_1$, les valeurs de $(\tilde{\beta}^1, g_1)$ sont fixées.

$$(\tilde{\beta}^2)^* = ((\mathbf{X}^2)'\mathbf{X}^2)^{-1}(\mathbf{X}^2)'\mathbf{X}^1\tilde{\beta}^1 \quad \text{et} \quad g_2^* = g_1\,.$$

Finalement, pour obtenir des modèles bayésiens de régression équitables :

1) utiliser la loi a priori de Zellner pour le modèle complet ;

2) en déduire les lois a priori des $2^p - 1$ modèles restants en prenant pour chaque modèle la loi a priori équitable par rapport au modèle complet.

$$\beta^\gamma | \mathbf{X}, \gamma, \sigma^2 \sim \mathcal{N}_{p_\gamma+1} \left(\left((\mathbf{X}^\gamma)' \mathbf{X}^\gamma \right)^{-1} \mathbf{X}^\gamma \mathbf{X} \tilde{\beta}, g\sigma^2 \left((\mathbf{X}^\gamma)' \mathbf{X}^\gamma \right)^{-1} \right) ,$$

($g = g_{(1,\ldots,1)}$ et $\tilde{\beta} = \tilde{\beta}^{(1,\ldots,1)}$).

Pour le paramètre γ, nous utilisons la loi a priori suivante

$$\pi(\gamma|\mathbf{X}) = \prod_{i=1}^{p} \tau_i^{\gamma_i}(1-\tau_i)^{1-\gamma_i},$$

où τ_i correspond à la probabilité a priori que la variable i soit présente dans le modèle.

Typiquement, lorsque aucune information a priori n'est présente, on pose $\tau_1 = \ldots = \tau_p = 1/2$.

Lois a priori non informatives

Dans un contexte non informatif, de nombreux auteurs ont proposé d'utiliser des lois a priori de Zellner centrées ($\tilde{\beta} = 0_{p+1}$) et de fixer une valeur pour g : $g = n$, $g = p^2$, $g = \max(n, p^2)$.

Le choix $g = n$ correspond à donner le poids d'une observation à la loi a priori. Pour n suffisamment grand, ce choix est proche du critère BIC.

Lorsque les lois a priori de Zellner sont centrées, l'hyper-paramètre g joue le rôle d'un paramètre de rétrécissement.

Aucune des solutions évoquées ci-dessus ne fait consensus.

Nous proposons d'utiliser des lois a priori de Zellner centrées et d'introduire une loi a priori hiérarchique diffuse sur g.

La loi a priori de Jeffeys pour le couple (σ^2, g) est égale à

$$\pi(\sigma^2, g|\mathbf{X}) \propto \sigma^{-2}(g+1)^{-1}\,.$$

Dans ce cas, la loi a posteriori de γ est telle que

$$\pi(\gamma|\mathbf{X}, \mathbf{y}) \propto \pi(\gamma|\mathbf{X}) \frac{{}_2F_1(n/2, 1; (p_\gamma+3)/2; \mathbf{y}'\mathbf{P}^\gamma\mathbf{y}/\mathbf{y}'\mathbf{y})}{p_\gamma+1}\,,$$

où ${}_2F_1$ est la fonction gaussienne hypergéométrique.

$$\mathbb{E}(\beta^\gamma|\mathbf{X},\gamma,\mathbf{y},g) = \left(\frac{g}{g+1}\right)\hat{\beta}^\gamma\,,$$

$$\mathbb{E}(\beta^\gamma|\mathbf{X},\gamma,\mathbf{y}) = \left(2\,\frac{{}_2F_1(n/2,2;(p_\gamma+3)/2+1;\mathbf{y}'\mathbf{P}_\gamma\mathbf{y}/\mathbf{y}'\mathbf{y})}{(p_\gamma+3)\,{}_2F_1(n/2,1;(p_\gamma+3)/2;\mathbf{y}'\mathbf{P}_\gamma\mathbf{y}/\mathbf{y}'\mathbf{y})}\right)\hat{\beta}^\gamma\,,$$

$$\hat{\mathbf{y}}_{\text{new}} = \mathbb{E}\left[\mathbf{y}_{\text{new}}|\mathbf{X}_{\text{new}},\mathbf{X},\mathbf{y}\right]$$

$$= \left(2\,\frac{\sum_{\gamma\in\Gamma}{}_2F_1(n/2,2;(p_\gamma+3)/2+1;\mathbf{y}'\mathbf{P}_\gamma\mathbf{y}/\mathbf{y}'\mathbf{y})/\left[(p_\gamma+1)(p_\gamma+3)\right]\mathbf{X}_{\text{new}}\hat{\beta}^\gamma}{\sum_{\gamma\in\Gamma}{}_2F_1(n/2,1;(p_\gamma+3)/2;\mathbf{y}'\mathbf{P}_\gamma\mathbf{y}/\mathbf{y}'\mathbf{y})/(p_\gamma+1)}\right).$$

Des expérimentations ont illustré la capacité de cette procédure à sélectionner un ensemble parcimonieux de variables.

Nous avons montré que cette procédure produit de meilleurs résultats que les critères classiques, notamment pour des échantillons de faibles tailles.

Approximation par échantillonnage de Gibbs

Lorsque le nombre de variables p est grand, typiquement $p > 25$, il est impossible de réaliser une sélection exhaustive.

Remarquons que

$$\pi(\gamma_i|\mathbf{X}, \mathbf{y}, \gamma_{-i}) \propto \pi(\gamma|\mathbf{X}, \mathbf{y}) \,.$$

Comme γ_i est binaire, la distribution conditionnelle $\pi(\gamma_i|\mathbf{X}, \mathbf{y}, \gamma_{-i})$ est obtenue par le calcul normalisé de $\pi(\gamma|\mathbf{X}, \mathbf{y})$ pour $\gamma_i = 0$ et $\gamma_i = 1$.

Échantillonneur de Gibbs

L'estimateur de $\pi(\gamma|\mathbf{X},\mathbf{y})$ déduit de l'échantillonnage de Gibbs est

$$\widehat{\pi(\gamma|\mathbf{X},\mathbf{y})}^{GIBBS} = \left(\frac{1}{T-T_0}\right) \sum_{t=T_0+1}^{T} \mathbb{I}_{\gamma}(\gamma^t),$$

et celui de $\mathbb{P}(\gamma_i = 1|\mathbf{X},\mathbf{y})$ s'écrit

$$\widehat{\mathbb{P}(\gamma_i = 1|\mathbf{X},\mathbf{y})}^{GIBBS} = \left(\frac{1}{T-T_0}\right) \sum_{t=T_0+1}^{T} \mathbb{I}_{\gamma_i}(\gamma_i^t).$$

Lorsque le nombre de régresseurs est inférieur à 100, l'échantillonneur de Gibbs donne des résultats très satisfaisants même s'ils sont fortement corrélés.