
Sélection de variables en régression

Jean-Michel MARIN

Institut de Mathématiques et Modélisation
Université Montpellier 2

*Travaux en collaboration avec Gilles Celeux, Mohammed El Anbari et
Christian Robert*

Paradigme bayésien paramétrique

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

$\boldsymbol{\theta} \in \Theta$ est un paramètre inconnu.

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

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

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

Choix bayésien de modèles

$$\mathcal{M}_1 : \mathbf{y} \sim f_1(\mathbf{y}|\boldsymbol{\theta}_1), \boldsymbol{\theta}_1 \in \Theta_1, \boldsymbol{\theta}_1 \sim \pi_1(\boldsymbol{\theta}_1).$$

$$\mathcal{M}_2 : \mathbf{y} \sim f_2(\mathbf{y}|\boldsymbol{\theta}_2), \boldsymbol{\theta}_2 \in \Theta_2, \boldsymbol{\theta}_2 \sim \pi_2(\boldsymbol{\theta}_2).$$

Munissons l'espace des modèles d'une loi de probabilité a priori :

$$\mathbb{P}(\mathcal{M}_1) \text{ et } \mathbb{P}(\mathcal{M}_2).$$

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

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

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 $\pi_1(\boldsymbol{\theta}_1)$ et $\pi_2(\boldsymbol{\theta}_2)$.

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 : $\mathbf{y}, \mathbf{x}_1, \dots, \mathbf{x}_p$.

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

$$\mathbf{y} | \mathbf{X}, \boldsymbol{\gamma}, \boldsymbol{\beta}^\boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_n(\mathbf{X}^\boldsymbol{\gamma} \boldsymbol{\beta}^\boldsymbol{\gamma}, \sigma^2 I_n),$$

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

Le paramètre σ^2 est commun à tous les modèles.

Objectif : déterminer le modèle le plus pertinent parmi les 2^p modèles en compétition, inférer sur le paramètre $\boldsymbol{\gamma}$.

Modèle alternatif :

$$\mathbf{y} | \mathbf{X}, \gamma, \alpha, \boldsymbol{\beta}_{\text{inv}}^\gamma, \sigma^2 \sim \mathcal{N}_n (\alpha \mathbf{1}_n + \mathbf{X}_{\text{inv}}^\gamma \boldsymbol{\beta}_{\text{inv}}^\gamma, \sigma^2 I_n) ,$$

- $\mathbf{X}_{\text{inv}}^\gamma$ la matrice dont les colonnes sont composées du vecteur des variables \mathbf{x}_i dont $\gamma_i = 1$ (cette matrice est centrée),
- $\alpha \in \mathbb{R}$, $\boldsymbol{\beta}_{\text{inv}}^\gamma \in \mathbb{R}^{p_\gamma}$ et $\sigma^2 \in \mathbb{R}_+^*$ sont les paramètres inconnus.

Les paramètres σ^2 et α sont communs à tous les modèles.

Pour un choix de loi a priori adéquat, ce modèle assure l'invariance de la procédure de sélection de variables à toute translation et tout changement d'échelle sur le vecteur \mathbf{y} .

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,

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

Pour le modèle alternatif,

$$\beta_{\text{inv}}^\gamma | \mathbf{X}, \gamma, \sigma^2 \sim \mathcal{N}_{p_\gamma}(\tilde{\beta}_{\text{inv}}^\gamma, g_\gamma \sigma^2 ((\mathbf{X}_{\text{inv}}^\gamma)' \mathbf{X}_{\text{inv}}^\gamma)^{-1}),$$
$$\pi(\alpha, \sigma^2 | \mathbf{X}, \gamma) \propto \sigma^{-2}.$$

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

et

$$\mathbb{E}(\alpha | \mathbf{X}, \gamma, \mathbf{y}) = \bar{y}.$$

Compatibilité de $\pi_1(\boldsymbol{\theta}_1)$ et $\pi_2(\boldsymbol{\theta}_2)$: information de Kullback-Leibler entre $f_1(\mathbf{y}) = \int_{\boldsymbol{\theta}_1} f_1(\mathbf{y}|\boldsymbol{\theta}_1)\pi_1(\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1$ et $f_2(\mathbf{y}) = \int_{\boldsymbol{\theta}_2} f_2(\mathbf{y}|\boldsymbol{\theta}_2)\pi_2(\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2$.

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

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

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

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

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

Nous avons les mêmes types de résultat pour le modèle alternatif.

Finalemment, 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.

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

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

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 $\tau_1 = \dots = \tau_p = 1/2$.

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 et \mathbf{P}^γ est le projecteur sur le sous-espace vectoriel engendré par les colonnes de \mathbf{X}^γ .

Dans le cas du modèle alternatif assurant l'invariance, on ne peut pas utiliser une loi a priori diffuse sur g !

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

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

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

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

Simulation studies

Example 1: Sparse uncorrelated case ($\rho = 0$)

$p = 10$, the components of \mathbf{x}_i ($i = 1, \dots, 10$) are generated independently as $\mathcal{N}_1(0, 1)$ realizations, and

$$\mathbf{y} \sim \mathcal{N}_n(2 + \mathbf{x}_2 + 2\mathbf{x}_3 - 2\mathbf{x}_6 - 1.5\mathbf{x}_7, I_n).$$

Example 2: Sparse correlated case ($\rho = 0.9$)

$$p = 10,$$

$$i = 1, 2, \quad \mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{11})/\sqrt{10},$$

$$i = 3, 4, 5, \quad \mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{12})/\sqrt{10},$$

$$i = 6, \dots, 10, \quad \mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{13})/\sqrt{10},$$

(the components of \mathbf{z}_i ($i = 1, \dots, 13$) are generated independently as $\mathcal{N}_1(0, 1)$ realizations), and

$$\mathbf{y} \sim \mathcal{N}_n(2 + \mathbf{x}_2 + 2\mathbf{x}_3 - 2\mathbf{x}_6 - 1.5\mathbf{x}_7, I_n).$$

$$\begin{bmatrix} 1 & 0.9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.9 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0.9 & 0.9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9 & 1 & 0.9 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.9 & 0.9 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0.9 & 0.9 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 1 & 0.9 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 1 & 0.9 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 0.9 & 1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0.9 & 0.9 & 0.9 & 0.9 & 1 \end{bmatrix} \cdot$$

Example 3: Sparse noisy correlated case

$p = 8$, the components of \mathbf{x}_i ($i = 1, \dots, 8$) are such that $\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}$
and

$$\mathbf{y} \sim \mathcal{N}_n(3\mathbf{x}_1 + 1.5\mathbf{x}_2 + 2\mathbf{x}_5, 9I_n).$$

Example 4: Not sparse noisy correlated case

Example 4 is the same as Example 3, except that $\beta_j = 0.85$ for all j
creating a non-sparse underlying model.

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BRIC	g prior with $g = \max(n, p^2)$
EB-L	Local EB estimate of g in g -prior
EB-G	Global EB estimate of g in g -prior
HG-3	Hyper- g prior with $a = 3$ Liang et al. (2008)
HG-4	Hyper- g prior with $a = 4$ Liang et al. (2008)
US	Our procedure
LASSO	Lasso Tibshirani (1996)
DZ	The Dantzig Selector Candes, E. and Tao, T. (2007)
ENET	The elastic-net Zou and Hastie (2005)

Table 1: Methods

Each data set consists of a training set V of size $n_V = 50$, on which the model where fitted and a test set T of size $n_T = 200$ for evaluation of the performance.

Tuning parameters in the Lasso, the Dantzig selector (DZ), the elastic-net (ENET) were selected by ten fold cross-validation.

For each example, we simulated 100 data sets.

	MSE_y	HITS	FP
AIC	1.088450	04	01
BIC	1.066091	04	00
BRIC	1.060175	04	00
EB-L	1.057609	04	00
EB-G	1.057667	04	00
HG-3	1.057877	04	00
HG-4	1.057877	04	00
US	1.056889	04	00
LASSO	1.104694	04	03
DZ	1.072656	04	00
ENET	1.112682	04	03

Table 2: Example 1: Sparse uncorrelated case

	MSE_y	HITS	FP
AIC	1.086974	04	01
BIC	1.071118	04	00
BRIC	1.057312	04	00
EB-L	1.049287	04	00
EB-G	1.049234	04	00
HG-3	1.049131	04	00
HG-4	1.049131	04	00
US	1.048992	04	00
LASSO	1.094448	04	04
DZ	1.122930	04	05
ENET	1.099196	04	04

Table 3: Example 2: Sparse correlated case

	MSE_y	HITS	FP
AIC	3.246930	03	01
BIC	3.209046	03	00
BRIC	3.207557	03	00
EB-L	3.213960	03	00
EB-G	3.215752	03	00
HG-3	3.213897	03	00
HG-4	3.213897	03	00
US	3.211109	03	00
LASSO	3.244892	03	02
DZ	3.169676	03	01
ENET	3.240923	03	02

Table 4: Example 3: Sparse noisy correlated case

	MSE_y	HITS	FP
AIC	3.421458	05	00
BIC	3.522257	04	00
BRIC	3.518709	04	00
EB-L	3.462775	04	00
EB-G	3.459188	04	00
HG-3	3.459278	04	00
HG-4	3.459278	04	00
US	3.470033	04	00
LASSO	3.259699	08	00
DZ	3.270812	07	00
ENET	3.234707	08	00

Table 5: Example 4: Not sparse noisy correlated case

Real dataset example: body fat dataset

The body fat dataset has been used by Penrose, Nelson and Fisher (1985).

The study aims at the estimation of the percentage of body fat by various body circumference measurements for 252 men.

The thirteen regressors are age (1), weight (lbs) (2), height (inches) (3), neck circumference (4), chest circumference (5), abdomen 2 circumference (6), hip circumference (7), thigh circumference (8), knee circumference (9), ankle circumference (10), biceps (extended) circumference (11), forearm circumference (12), and wrist circumference (13).

In order to investigate the performances, the data set has been split 25 times into a training set of 151 observations and a test set of 101 observations.

Tuning parameter have been chosen by tenfold cross validation.

	MSE_y	Median number of selected variables
AIC	4.649241	07
BIC	4.619032	04
BRIC	4.680235	05
EB-L	4.636144	04
EB-G	4.677880	04
HG-3	4.671588	04
HG-4	4.678505	04
US	4.636659	02
LASSO	4.631979	11
DZ	4.632179	12
ENET	4.600197	11

Table 6: Body fat dataset example

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
AIC	13	18	05	14	02	25	10	08	02	04	08	20	21
BIC	01	20	02	04	00	25	03	02	00	01	02	08	17
BRIC	01	20	02	04	00	25	04	02	01	03	10	07	18
EB-L	01	20	02	04	00	25	03	02	00	01	03	08	16
EB-G	01	20	02	04	00	25	04	02	01	01	03	10	18
HG-3	01	20	02	04	00	25	03	02	00	01	03	08	16
HG-4	01	20	02	04	00	25	04	02	00	01	03	10	18
US	00	19	01	01	00	25	01	01	00	00	00	02	09
LASSO	25	15	24	24	13	25	16	17	13	18	16	22	25
DZ	25	23	25	24	16	25	23	20	11	24	23	25	25
ENET	25	14	25	22	12	25	17	18	14	22	18	23	25

Table 7: Body fat dataset example: the number of times a variable has been selected into the 25 random splits

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

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

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.