PROBABILISTIC GRAPHICAL MODEL

Guillaume PETIT
Thomas FAURÉ

1 Question 1

La normalisation permet de mettre à la même échelle toutes les données. Cela est utile quand on sait que la loi à priori des β à une variance constante. Ajouter un intercepte permet de passer d'un modèle linéaire à un modèle affine.

2 Question 2

Soit $\hat{\epsilon}_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Alors

$$y_i = sign(\beta^T x_i + \hat{\epsilon}_i) = sign(\beta^T x_i + \sigma\epsilon_i) = sign(\hat{\beta}^T x_i + \epsilon_i)$$

avec $\hat{\beta} = \beta/\sigma$

3 Question 3

On cherche la loi de $\beta, z|X, y$. Selon l'article "Gibbs Sampling for the Probit Regression Model with Gaussian Markov Random Field Latent Variables", l'échantillonneur de Gibbs implique un échantillonnage itératif de $\beta|z$ et de $z|\beta, y$.

•$\beta|z$: Par définition, $z|\beta \sim \mathcal{N}(X\beta, I)$

$$\begin{aligned} p(\beta|z) &\propto p(\beta)p(z|\beta) \\ &\propto exp\left[-\frac{1}{2}\left(||z - X\beta||^2 + \frac{1}{\tau}\beta^T\beta\right)\right] \\ &\propto exp\left[-\frac{1}{2}(\beta - \mu_p)^T\Sigma_p^{-1}(\beta - \mu_p)\right] \end{aligned}$$

avec $\mu_p = \Sigma_p X^T z$ et $\Sigma_p^{-1} = X^T X + \frac{1}{\tau} I_p$. Donc $\beta|z \sim \mathcal{N}(\mu_p, \Sigma_p)$

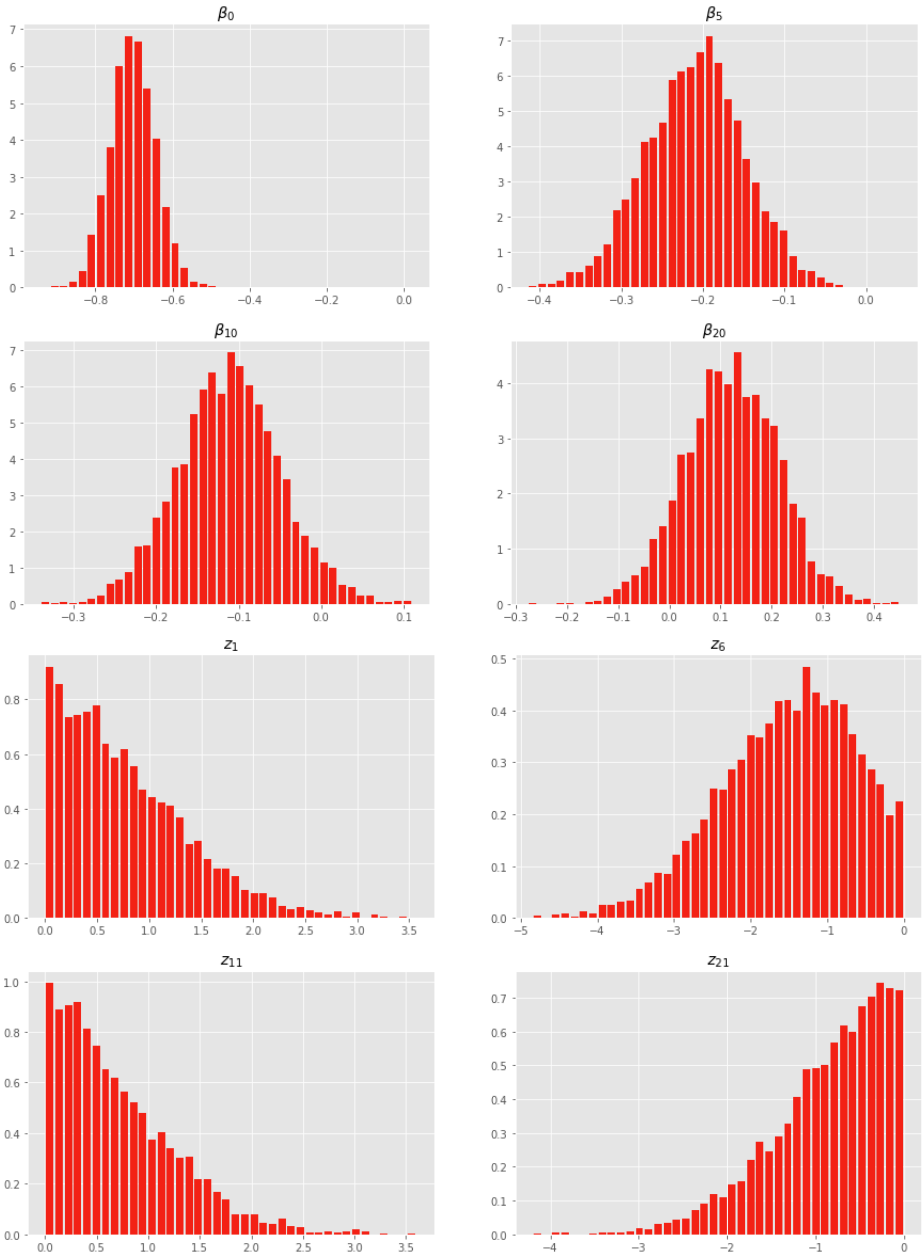
•$z|\beta, y$:

$$p(z|\beta, y) \propto p(z, \beta, y) \propto p(z|\beta)p(y|z, \beta) = \prod_{i=1}^{N} p(z_i|\beta)p(y_i|z_i)$$

Alors, $p(z_i|\beta, y_i) \propto \begin{cases} \mathcal{N}(\beta^T x_i, 1)\mathcal{I}(z_i > 0, y_i = 1) \\ \mathcal{N}(\beta^T x_i, 1)\mathcal{I}(z_i < 0, y_i = -1) \end{cases}$

avec $\mathcal{I}(A)$ l'indicatrice de l'évènement A.

On divise le jeu de données en partie training (80 %) et test (20/%). On inférence les variables β et z par l'algorithme de Gibbs sur la partie training. On obtient la distribution suivante pour ces paramètres :



Titre : distribution de la loi de $\beta, z|X, y$ avec l'algorithme de Gibbs

Une fois les variables observées, nous prédisons les labels pour la partie test. Pour la prédiction, on sait que la fonction minimisant la perte est

$$g^*(X,\beta) = 2\mathcal{I}(\eta(X,\beta) > 0.5) - 1$$

avec $\eta(X,\beta) = \mathbb{P}(y=1|X,\beta) = \Phi(X\beta)$ avec Φ la fonction de répartition d'une loi normale centrée réduite. On estime η selon la loi des grands nombres.

4 Question 4

Nous cherchons à approcher la loi de $p(\beta, z|X, y)$ par une loi $q(\beta, z)$ qui vérifie $q(\beta, z) = q_1(\beta)q_2(z)$.
D'après le cours nous savons que :

$$q_1(\beta) \propto exp\left\{\mathbb{E}_{z\sim q_2}\left[log\ p(\beta|z,y)\right]\right\}$$

et

$$q_2(z) \propto exp\left\{\mathbb{E}_{\beta\sim q_1}\left[log\ p(z|\beta,y))\right]\right\}$$

Nous allons essayer de déterminer la loi de q_1 et q_2 en regardant $\mathbb{E}_{z\sim q_2}\left[log\ p(\beta|z,y)\right]$ et $\mathbb{E}_{\beta\sim q_1}\left[log\ p(z|\beta,y))\right]$ respectivement.

•Pour q_1 :

$$\mathbb{E}_{z\sim q_2}\left[log\ p(\beta|z,y)\right]$$

$$= \mathbb{E}_{z\sim q_2}\left[log\ p(\beta|z)\right]$$

$$= \mathbb{E}_{z\sim q_2}\left[-\frac{1}{2}[(\beta-\mu_p)^T\Sigma_p^{-1}(\beta-\mu_p)]\right] + cte$$

$$= -\frac{1}{2}\left[(\beta-\overline{\mu_p})^T\Sigma_p^{-1}(\beta-\overline{\mu_p})\right] + cte$$

avec $\overline{\mu_p} = \Sigma_p X^T \mathbb{E}_{z\sim q_2}(z)$.

Finalement on a : $q_1(\beta) \propto exp\left\{-\frac{1}{2}\left[(\beta-\overline{\mu_p})^T\Sigma_p^{-1}(\beta-\overline{\mu_p})\right]\right\}$,
on en conclut que $q_1(\beta) = \mathcal{N}(\overline{\mu_p}, \Sigma_p)$

•Pour q_2 :

$$\mathbb{E}_{\beta\sim q_1}\left[log\ p(z|\beta,y))\right]$$

$$= \mathbb{E}_{\beta \sim q_1} \left[-\frac{1}{2}||z - X\beta||^2 + \sum_{i=1}^{n} log\ \mathcal{I}(z_i y_i > 0) \right] + cte$$

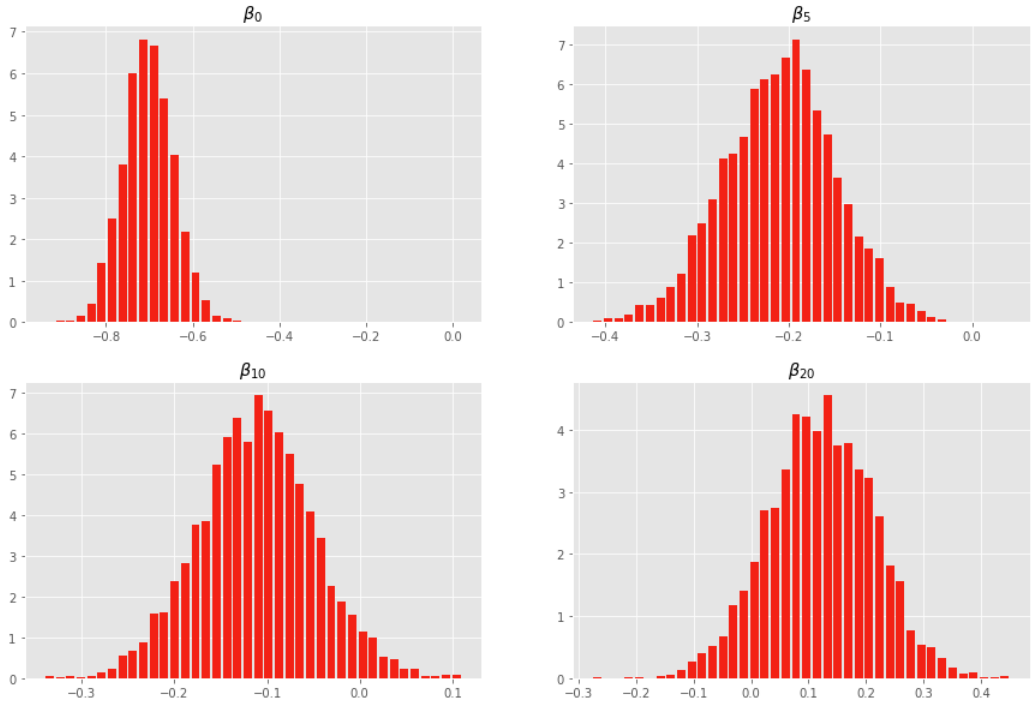
$$= -\frac{1}{2}||z - X\mathbb{E}_{\beta \sim q_1}(\beta)||^2 + \sum_{i=1}^{n} log\ \mathcal{I}(z_i y_i > 0) + cte$$

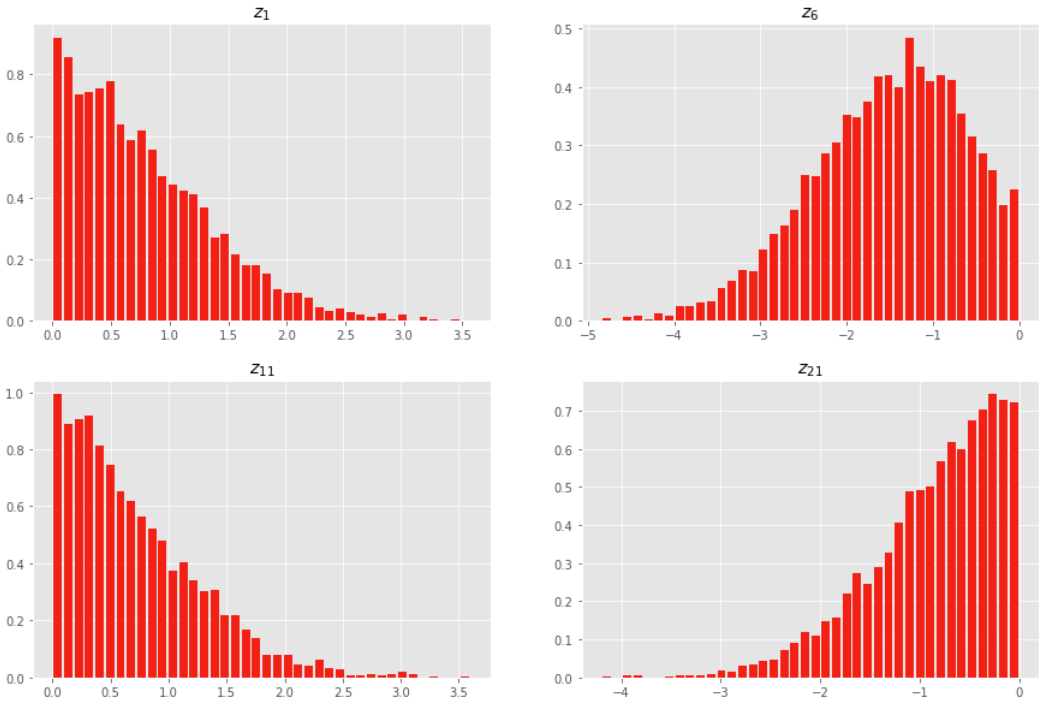
Finalement on a : $q_2(z) \propto exp\left\{-\frac{1}{2}||z - X\mathbb{E}_{\beta \sim q_1}(\beta)||^2 + \sum_{i=1}^{n} log\ \mathcal{I}(z_i y_i > 0)\right\}$
on en conclut que $q_2 = \mathcal{N}^{P_y}(X\mathbb{E}_{\beta \sim q_1}[\beta], I)$ où
$P_y = \{z \in \mathbb{R}^n | z_i y_i > 0, \forall i \in [1, n]\}$ et $\mathcal{N}^g(\mu, \Sigma)$ une loi normale tronquée sur l'ensemble G.

Résultats :

•**Accuracy** : Gibbs et Mean Field ont des résultats à peu près similaires avec une meilleure performance pour Gibbs. En moyenne, Gibbs fournit 76 % de précision tandis que MF est plutôt aux alentours de 74%.

•**Vitesse** : Pour 4000 itérations, Gibbs prend en moyenne 30 secondes tandis que MF prend 40 secondes. (Les résultats auraient pu être réduits de moitié en travaillant avec des array au lieu de dataframe, mais les calculs donnaient des résultats erronés pour une raison inconnue).

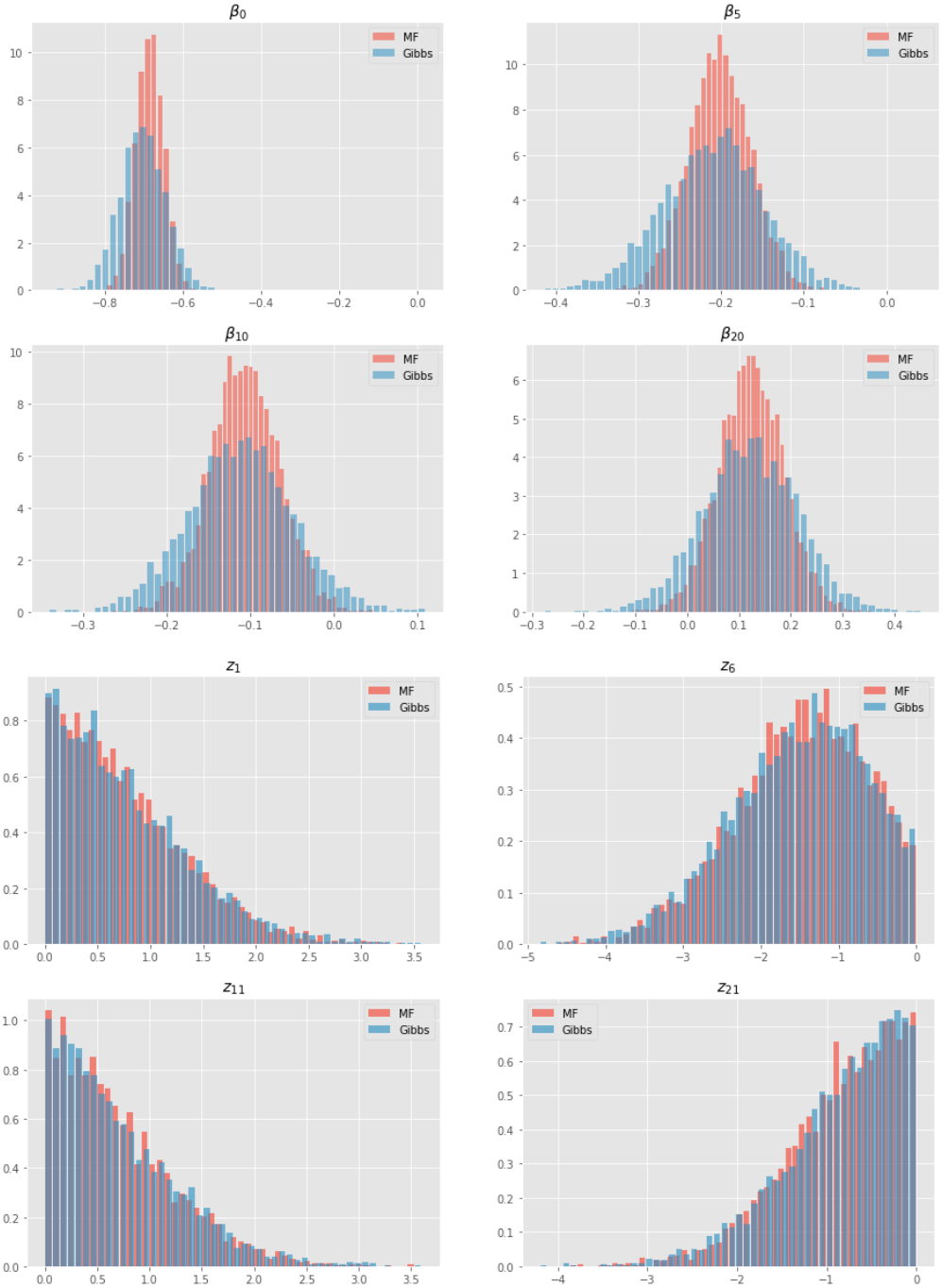




Titre : distribution de la loi de $\beta, z|X, y$ avec l'algorithme Mean Field

5 Question 5

L'algorithme Mean Field repose sur la minimisation de l'entropie relative KL. Or dans ce cas, KL est une espérance sous la loi $q(\beta, z)$. Ainsi, cet algorithme va accorder un poids plus important sur les tirages centrés que sur les tirages aberrants. Cette pondération va donc réduire fortement la variance et avoir une distribution plus importante au niveau de sa moyenne. Ce qu'on peut d'ailleurs observer en superposant les deux distributions :



Comparaison de la distribution de la loi $\beta, z|X, y$ selon les algorithmes

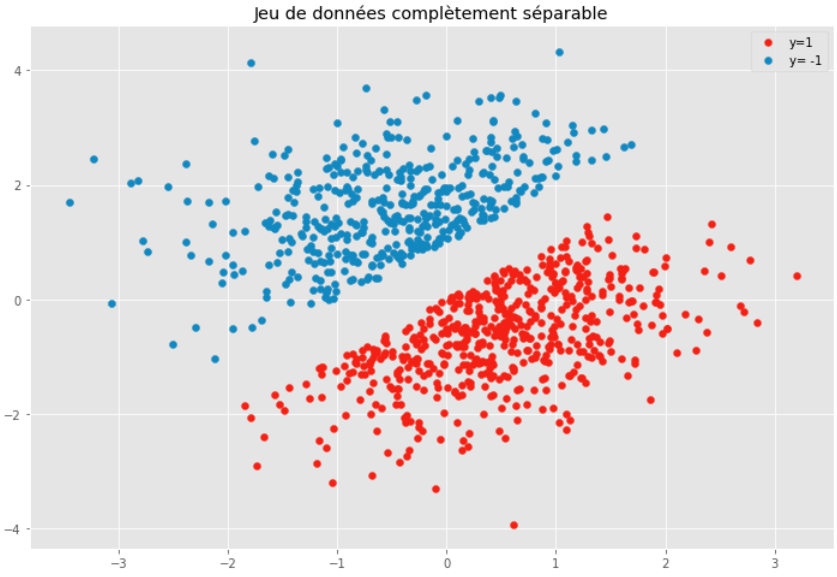
6 Question 6

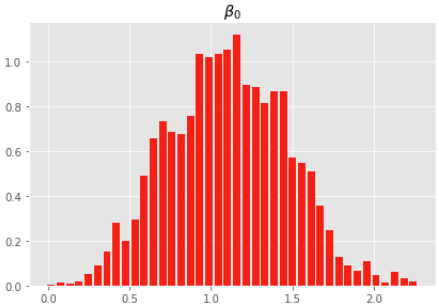
La "meilleure" droite séparant les données est celle qui a le meilleur compromis entre les points mal classés et ceux bien classés. Dans le cas séparation parfaite, on $\forall \beta \in \mathbb{R}^p, y_i \beta^T x_i \geq 0$. Ainsi, si on multiplie par $K \in \mathbb{R}$ tous les β, cela ne changerait pas l'équation de la droite et ne pénaliseraient pas les mauvaises classification car il n'y en a pas. Ainsi, dans certains cas comme dans la régression logistique ou probit, le maximum de vraisemblance peut valoir $+\infty$.

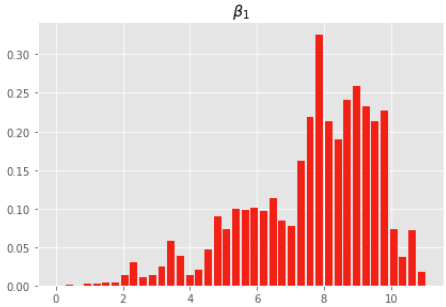
Nous créons notre jeu de données complètement séparable (voir ci-dessous) par la droite d'équation $x_1 - x_2 + 0.5 = 0$. Après avoir fait tourner l'algorithme de Gibbs, on remarque que nos prédictions sont parfaites bien que nos β à posteriori ne suivent pas du tout les lois voulues. Il n'y a sûrement pas eu convergence. Or, on a vu dans le cours que

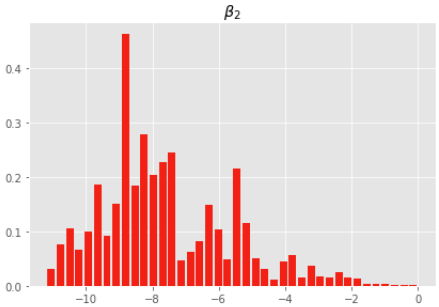
$$\beta_{MAP} = \underset{\beta}{argmax} \left\{ log\ p(y|\beta) - \frac{\tau}{2}||\beta||^2 \right\}$$

On peut voir le τ comme le coefficient de pénalisation d'une régression Ridge. Ainsi, on a tenté d'observer les résultats qu'on pouvait obtenir en augmentant cette pénalisation. Les résultats montrent que les coefficients finissent bien par converger vers une loi normale. Néanmoins, les valeurs que prennent β_0 ont une grande variance et ne sont pas aux alentours de 0.5, ce qui impactent les valeurs de β_2 qui sont donc un peu plus négatives que voulues, expliquant la précision de 0.95 %. Il faut toutefois noter que cette pénalisation n'est pas toujours effective.

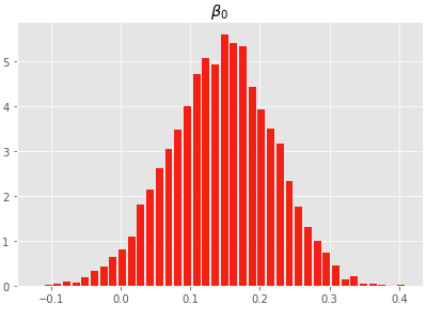


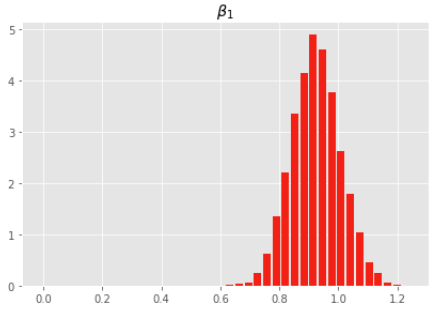


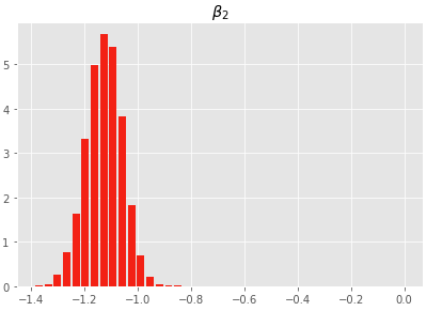




Distribution des β avec une petite pénalisation ($\tau = 100$)







Distribution des β avec une grande pénalisation ($\tau = 1000$)