setting:

$y = \beta_1 x_1 + \cdots + \beta_p x_p + \sigma \varepsilon$

for $\varepsilon \sim N(0,1)$

parameters $\beta \in \mathbb{R}^p$, $\sigma > 0$

n observations

$(x_1, y_1) \cdots (x_n, y_n)$

But $p >>> n$

(more parameters than observations)

If you use OLS, it is not well-defined

Assumptions: β is sparse,
In other words, a lot of β's are zero, and we need to have this assumption to have identifiability.

From Bayesian perspective:

where is the uncertainty?

① number of β's with entries 0 ?
② which β's are zero?
③ the value of β's with non-zero entry?

Encoding the belief for ③
By introducing

$\gamma_1, \gamma_2 \dots \gamma_p \in \{0, 1\}$

$\gamma_j \in 1 \Leftrightarrow \beta_j \neq 0 \Leftrightarrow$ "x_j is important"

$\gamma_j \in 0 \Leftrightarrow \beta_j = 0 \Leftrightarrow$ "x_j is not important"

$\Leftrightarrow$ means if and only if

$$\beta_j | r_j \sim \delta_0(1-r_j) + r_j \, \text{Unif}[-T, T]$$

for T large.

that is to say, if $r_j = 0$, then it is just a point mass 0.

if $r_j = 1$, you know β_j not zero, then you impose a super flat prior on it.

for the belief for ②

$r_1, r_2 \ldots r_p \mid \theta \sim \text{Bernoulli}(\theta)$

for the belief for ①

$\theta \sim \text{Beta}(a, b)$, and you would choose a and b to give small expectation of θ as you believe β's are sparse.

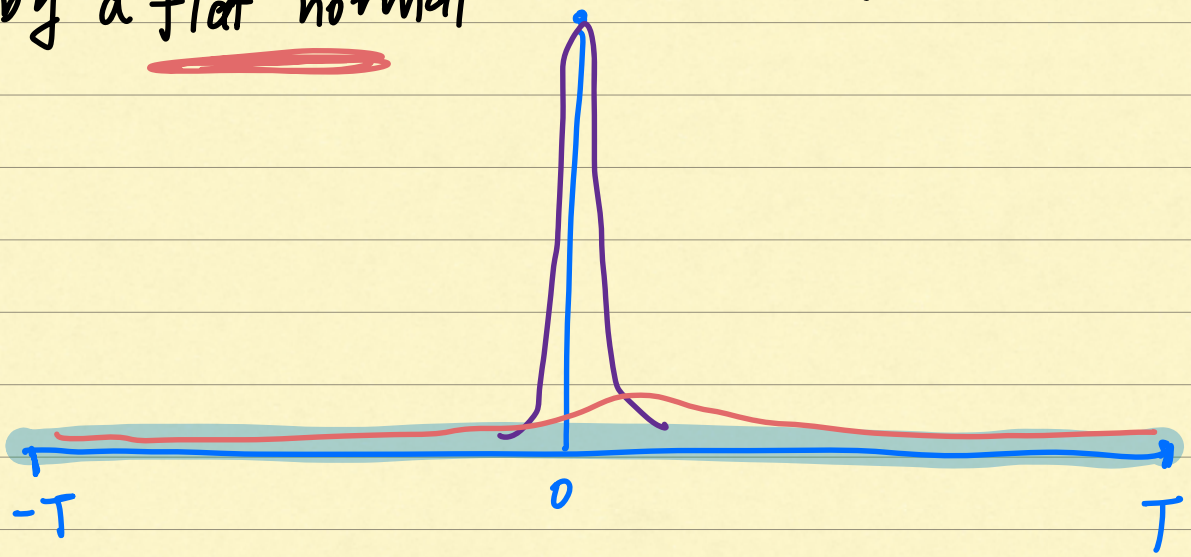
However, using MCMC is not feasible

Challenge!

as $r_1, r_2 \ldots r_p \mid \theta$ needs updating at the same time and you have too many configurations.

challeuge 2 : the graph is not differentiable

continuous relaxation, a modification of the graph, replace the point mass by a peaky normal, and uniform replaced by a flat normal



your model becomes

$$\theta \sim \text{Beta}(a, b)$$

$$r_1 \dots r_p | \theta \sim \text{Bernoulli}(\theta)$$

$$\beta_j | r_j, \sigma^2 \sim r_j \cdot N(0, v_1\sigma^2) + (1-r_j) N(0, v_0\sigma^2)$$

$$v_1 >>>>> v_0$$

challenge 1:

Not to know for sure the posterior density for each β, but just find the MAP.

$$\text{argmax}\ p(\beta, \theta, \sigma^2 \mid y)$$

$$p(\beta \mid \sigma^2, \theta) = \prod_{j=1}^{p} [\theta \cdot \text{slab density} + (1-\theta) \cdot \text{spike density}]$$

But the addition is hard to work with because if you take log. doesn't factorize.

$$\log \prod_{j=1}^{p} [\theta \cdot \text{slab density} + (1-\theta) \cdot \text{spike density}]$$

$$= \sum \log\left(\theta \frac{1}{\sqrt{2\pi\tau_1\sigma^2}} e^{-\beta_j^2/2\tau_1\sigma^2} + (1-\theta) \frac{1}{\sqrt{2\pi\tau_0\sigma^2}} e^{-\beta_j^2/2\tau_0\sigma^2}\right)$$

↓

Hard to optimize,

So you replace that with its lower bound

$$\mathbb{E}_{r_i} \log p(\beta, \theta, \sigma^2, r \mid y)$$

You know

$$\log p(\beta, \theta, \sigma^2, r \mid y) \geq \mathbb{E}_{r_i} \log p(\beta, \theta, \sigma^2, r \mid y)$$

By Jensen's inequality

thus, for each iteration

- optimize the $\mathbb{E}_{r_i} \log p(\beta, \theta, \sigma^2, r \mid y)$ to get new $\beta, \theta, \sigma^2, r$
- take Expectation

$$p(r \mid \beta, \theta, \sigma^2, y) \propto \prod_{j=1}^{p} \left(\frac{1}{2v_1\sigma^2} e^{-\beta_j^2 / 2v_1\sigma^2} \right)^{r_j} \left(\frac{1}{2v_0\sigma^2} e^{-\beta_j^2 / 2v_0\sigma^2} \right)^{1-r_j}$$

$p_1(\beta_j)$ $\qquad$ $p_0(\beta_j)$

$$\cdot \prod_{j=1}^{p} \theta^{r_j} (1-\theta)^{1-r_j}$$

thus

$$p(r_j=1 \mid \ ) \propto \theta \cdot P_1(B_j)$$

$$p(r_j=0 \mid \ ) \propto (1-\theta) \cdot P_0(B_j)$$

define $p_j^* \equiv \dfrac{\theta P_1(B_j)}{\theta P_1(B_j) + (1-\theta) P_0(B_j)}$

write out $p(\beta, \theta, \sigma^2, r \mid y)$

$$-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\Sigma(y_i - x_i^T\beta)^2$$

$$+\Sigma\left[r_j \log P_1(B_j) + (1-r_j)\log P_0(B_j)\right]$$

$$+\Sigma[r_j \log\theta + (1-r_j)\log(1-\theta)$$

$$+\log p(\theta) + \log p(\sigma^2)$$

and $r \mid \beta, \sigma^2, \theta, y$ would be

$$-\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum(y_i - x_i^T\beta)^2$$
$$+\sum\left[p_j^* \log p_1(\beta_j) + (1-p_j^*)\log p_0(\beta_j)\right]$$
$$+\sum p_j^* \log\theta + (1-p_j^*)\log(1-\theta)$$
$$+\log p(\theta) + \log p(\sigma^2)$$

So update that via EM.

$$r \mid \beta^{(t)}, \sigma^{2(t)}, \theta^{(t)}, y$$

and recall,

$$p_1(\beta_j) = (2\pi\sigma^2 \cdot v_1)^{-1/2} e^{-\beta_j^2/2v_1\sigma^2}$$

thus

$$\log p_1(\beta_j) = \frac{-\beta_j^2}{2v_1\sigma^2} - \frac{1}{2}\log\sigma^2$$

so for the optimization step,

(EMVS

the objective function is

$$\Sigma (y_i - x_i^T \beta)^2 - \frac{1}{2} \Sigma \beta_j^2 \left[\frac{p_j^*}{v_1} + \frac{1-p_j^*}{v_0} \right]$$

It looks like a Ridge

But the penalty depends if you come from spike, small penalty; if you come from slab, Big penalty.