# BSM4 - Model assessment and improvement

Lecturer: Riccardo Corradin

# Introduction

This slide block presents some of the fundamental strategies to **assess** a specific estimated regression model, **compare** the model with other competitors and **improve** the model by slightly changing its specification.

We will tackle mainly the following methodologies.

- How to **assess** and **evaluate the goodness of fit** once we have a posterior distribution, or a sample from such a distribution.
  - → Tracing the empirical evidence.
  - → Constructing indices based on the empirical evidence.
- How to **compare two distinct models**, in terms of model fit or other relevant measures.
  - → Fit measures and tests.
- How to **select** a set of **relevant covariates** for a particular model specification.
  - → Scanning the whole model space.
  - → Spike-and-slab priors
- How to **relax the model specification** and being more vague on the prior guess, for example resorting to deeper hierarchies in the model setting.

Most of the methodologies here presented are not restricted to regression models we saw in the last week. However, examples will be related to them.

As before, we assume that we have $y_{1:n} \in \mathbb{Y} \subseteq \mathbb{R}$ response variables that we want to model as function of $\boldsymbol{x}_{1:n} \in \mathbb{X} \subseteq \mathbb{R}^d$ suitable transformations of covariates.

# Model assessment and comparisons

# Model assessment and comparison

From a **frequentist perspective**, we know that if we evaluate a density function of an observation $y_i$ at a specific value of a regression parameter $\boldsymbol{\theta}$ and possibly other parameters,

$$f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}),$$

we are measuring **how likely** is $y_i$ given the probabilistic model assumption we are taking $f(\cdot \mid \cdot)$ and the parameter values $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \dots\}$, with $\boldsymbol{\theta} \in \Theta$.

→ The higher, the better.

Combining more observations, we construct our dear log-likelihood function

$$\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}).$$

At each term of the product in the right-hand side expression, the same of above applies.

→ Most fit measures and information criteria in the literature are based on likelihood function transformations, evaluated at a particular point estimate.

→ For historical reasons, predictive accuracy measures are called information criteria.

→ Typically, they are defined as function of the deviance, i.e. $-2 \log \mathcal{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \theta)$.

---

But what can we do in a **Bayesian framework**, where the **parameter** of interest is a **random quantity**?

# Model assessment and comparison

Here, we can resort to something that we already saw. We start from the predictive distribution of our model. For a new observation $\tilde{y}$, with covariates $\tilde{\boldsymbol{x}}$, the **posterior predictive distribution** is given by

$$m(\tilde{y} \mid \tilde{\boldsymbol{x}}, y_{1:n}, \boldsymbol{x}_{1:n}) = \int_{\Theta} f(\tilde{y} \mid \tilde{\boldsymbol{x}}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \mathrm{d}\boldsymbol{\theta},$$

with $\pi(\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n})$ being the posterior distribution given the observed sample.

→ Considering the previous, we do not have anymore the problem of having a random parameter.

Ideally, we can resort to the log-pointwise predictive density

$$\mathrm{LPPD} = \sum_{i=1}^{n} \log m(y_i \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \sum_{i=1}^{n} \log \left[ \int_{\Theta} f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \pi(\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \mathrm{d}\boldsymbol{\theta} \right].$$

to measure how much likely are the data under our model setting, once we marginalize the parameters.

- We are looking at the **predictive performance** of the estimated model.
- New problem, the posterior predictive in the previous depends on the whole data
  - → We use the generic $i$th observation to both estimate the model and compute the LPPD.

# Model assessment and comparison

**Several different criteria** are available in the literature. All of them are somehow approximating the out-of-sample predictive performance. Hence, all of them have flaws, but we need to measure the performance of our models.

We distinguish mainly among three families.

- **Within sample predictive accuracy**. A rough estimate of the expected predictive distribution of new data is given by what happens with the observed data, such as the LPPD. In general, quick to evaluate and easy to interpret.
- **Adjusted within sample predictive accuracy**. Quantities such as AIC, DIC, WAIC are adjusted measure by a model-complexity term.
- **Cross-validation**. Ideally, we separate our data in train and test, then we evaluate predictive performance on a subset of data.

In the following, we present some of the fundamental measures used to evaluate a model fit.

# Model assessment and comparison

- **Akaike information criteria (AIC)**. Usually, inference on a parameter of interest, say $\boldsymbol{\theta}$, is summarized by a point estimate $\hat{\boldsymbol{\theta}}$, typically the maximum likelihood estimate. The AIC is defined as

$$AIC = -2\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \hat{\boldsymbol{\theta}}_{ML}) + 2p.$$

- Ideally, ignoring the multiplication by $-2$, we are adjusting the log-likelihood by $p$, the number of parameters in our model. Such a quantity mitigates the fact that as far as we are increasing more parameters in the model, the predictive performance increases as well.
- In practice, the correction term adjust for overfitting.
- $p$ plays the role of effective number of parameters.
- As far as we are departing from a linear model with flat prior assumptions, we cannot simply add $p$.
- **The smaller, the better**.

# Model assessment and comparison

- **Deviance information criteria (DIC)**. Somehow, it is a Bayesian extension of AIC. We replace the maximum likelihood estimate $\hat{\boldsymbol{\theta}}_{ML}$ with the posterior mean of the parameter $\mathrm{E}[\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}]$, and the complexity penalization term $p$ with a data-based correction. Hence, we have

  $$DIC = -2\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \mathrm{E}[\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}]) + 2p_{DIC},$$

  where

  $$p_{DIC} = 2\left[\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \mathrm{E}[\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}]) - \mathrm{E}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}}[\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta})]\right]$$

  or

  $$\begin{aligned} p_{\mathrm{DIC}} &= 2\mathrm{var}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}}\left(\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta})\right) \\ &\approx \frac{2}{R-1}\sum_{r=1}^{R}\left[\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta}^{(r)}) - \overline{\ell(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\theta})}\right] \end{aligned}$$

- **The smaller, the better**.
- The variance term acts as a complexity penalization.
- For linear models with uniform priors, $p_{DIC}$ reduces to $k$.

# Model assessment and comparison

- **Widely applicable information criteria (WAIC)**. In practice, it is defined as a penalized version of $\mathrm{LPPD}$, i.e.,

$$\mathrm{WAIC} = -2\mathrm{LPPD} + 2p_{\mathrm{WAIC}},$$

where, given a MCMC sample $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(R)}\}$ from the posterior distribution

$$\mathrm{LPPD} = \sum_{i=1}^{n} \log m(y_i \mid y_{1:n}, \boldsymbol{x}_{1:n}) \approx \sum_{i=1}^{n} \log \left[ \frac{1}{R} \sum_{r=1}^{R} f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)}) \right],$$

and

$$p_{WAIC} = 2 \sum_{i=1}^{n} \left[ \log \mathrm{E}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}} \left[ f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \right] - \mathrm{E}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}} \left[ \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \right] \right]$$

or

$$\begin{aligned} p_{\mathrm{WAIC}} &= \sum_{i=1}^{n} \mathrm{var}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}} \left( \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}) \right) \\ &\approx \sum_{i=1}^{n} \frac{1}{R-1} \sum_{r=1}^{R} \left[ \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)}) - \overline{\log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right] \end{aligned}$$

with $\overline{\log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} = \frac{1}{R} \sum_{r=1}^{R} \log f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)})$.

- **The smaller, the better**.
- The variance term acts as a complexity penalization.
- For large $n$, $\mathrm{LPML} \approx -\frac{1}{2}\mathrm{WAIC}$.

# Model assessment and comparison

In the previous slide, we have been optimist by **including**, in each term, the same **observation in both the argument** of the density function and **the conditioning quantities**.

- Including the observation in the conditioning arguments **may alter the prediction** we are making, especially with small sample sizes.

Instead of the $\mathrm{LPPD}$, we can consider another quantity by replacing $m(y_i \mid y_{1:n}, \boldsymbol{x}_{1:n})$ with

$$m(y_i \mid y_{-i}, \boldsymbol{x}_{-i}) = \int_\Theta f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})\pi(\boldsymbol{\theta} \mid y_{-i}, \boldsymbol{x}_{-i})\mathrm{d}\boldsymbol{\theta},$$

where $y_{-i}$ and $\boldsymbol{x}_{-i}$ denote $y_{1:n}$ and $\boldsymbol{x}_{1:n}$ but discarding the $i$th element, respectively.

→ Ideally, we want to perform a leave-one-out cross-validation.

We then resort to the so-called log pseudo-marginal likelihood (LPML), which is given by

$$\mathrm{LPML} = \sum_{i=1}^{n} \log m(y_i \mid \boldsymbol{x}_i, y_{-i}, \boldsymbol{x}_{-i}) = \sum_{i=1}^{n} \log(\mathrm{CPO}_i).$$

where $\mathrm{CPO}_i$ is called the **conditional predictive ordinate** for observation $i$.

→ We have a new problem, we have to estimate $n$ separate models, one for each observation removed. Or maybe not...

# Model assessment and comparison

We can see that the generic $\mathrm{CPO}_i$ can be rewritten as

# Model assessment and comparison

# Model assessment and comparison

Hence, suppose we have a **MCMC sample** $\{\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(R)}\}$ from a **generic algorithm**. We can approximate each $\mathrm{CPO}_i$ as

$$\mathrm{CPO}_i = \mathrm{E}_{\boldsymbol{\theta} \mid y_{1:n}, \boldsymbol{x}_{1:n}} \left[ \frac{1}{f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta})} \right]^{-1} \approx \left[ \frac{1}{R} \sum_{r=1}^{R} \frac{1}{f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)})} \right]^{-1},$$

and then the LPML becomes

$$\mathrm{LPML} \approx \sum_{i=1}^{n} \log \left\{ \left[ \frac{1}{R} \sum_{r=1}^{R} \frac{1}{f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}^{(r)})} \right]^{-1} \right\}.$$

- With the previous expression, we can easily compute the LPML by tracing by passing as output of our algorithm also the density evaluation for each observation and each iteration.
  - → In STAN we just need to include an extra line of code.
- The LPML can be used **raw as fit measure**, or **transformed** for example introducing a model complexity penalization.

# Model assessment and comparison

## Example

Let us consider the following response variable and covariates

```
set.seed(123); betatrue <- c(-2, 2);
gammatrue <- rbind(c(-1, 2, 4), c(3, -2, -4))
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
z3 <- round(rnorm(100, 0, 1), digits = 1)
z4 <- round(rnorm(100, 0, 1), digits = 1)
c <- rep(c(1, 2), each = 50)
X1 <- cbind(z1, z2)
U1 <- cbind(rep(1, 100), z3, z4)
X2 <- cbind(rep(1, 100), z1, z2)
U2 <- cbind(z3, z4)
tempmeans <- as.vector(X1 %*% betatrue) +
apply(cbind(U1,gammatrue[c,]), 1, function(x) x[1:2] %*% x[3:4])
y <- sapply(tempmeans, function(x) rnorm(1, x, 1))
```

Consider the following two linear regression mixed model

$$LMM1: \qquad y_i = \beta_1 z_{i,1} + \beta_2 z_{i,2} + \gamma_{c_i,1} + \gamma_{c_i,2} z_{i,3} + \gamma_{c_i,3} z_{i,4} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2),$$

$$LMM2: \qquad y_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2} + \gamma_{c_i,1} z_{i,3} + \gamma_{c_i,2} z_{i,4} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2).$$

# Model assessment and comparison

LMM1 has a group-specific intercept term while LMM2 has a shared intercept term.

- Make a variation of the LMM gibbs sampler of slide block 3 to return as output, for each iteration, also the density evaluation of each observation, given the current value of the parameter.
- Produce a sample from the posterior distribution of the parameters for LMM1 and LMM2 and check mixing and convergence of the algorithm, assuming a priori $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathrm{diag}(10^3))$, $\gamma_j \sim N(\mathbf{0}, \mathrm{diag}(10^3))$, $j = 1, 2$, and $\sigma^2 \sim IG(2, 2)$.
- Write the functions to calculate LPML and WAIC. Evaluate the fit measures for both the estimated model.

Consider now as third model a linear regression model

$$LM1: \qquad y_i = \beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2} + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2).$$

- Write a function sampling from the posterior distribution of LM1, assuming $\boldsymbol{\beta} \sim N(\mathbf{0}, \mathrm{diag}(10^3))$ and $\sigma^2 \sim IG(2, 2)$.
- Evaluate the fit measures $\mathrm{LPML}$ and $\mathrm{WAIC}$. Compare them with the ones obtained with LMM1 and LMM2.
- Repeat the previous two points implementing all the models in STAN.

# Model assessment and comparison

Once we estimate **more than one model**, as commonly done in data analysis, we may be interested into selecting the **best** model, with respect to some specific criteria.

As general rule in statistical modelling, a model should be not only good in terms of fit and predictive performance, but also as simple as possible.

$\rightarrow$ Simple model are easy to interpret and explain.

One possible strategies is to calculate for each model quantities like $\mathrm{LPML}$ and $\mathrm{WAIC}$, then **rank the model best to worst**.

$\rightarrow$ However, even if we have a glimpse on possible predictive performance improvements, we do not see if two model significant differ from each other.

Suppose we have two distinct model, say $M_1$ and $M_2$. One first approach to select a model can be to perform comparison through tests. Hence, in a quite generic form, we have

$$\begin{aligned} M_1: &\qquad Y_i \sim f_1(y_i \mid \boldsymbol{x}_{1,i}, \boldsymbol{\theta}_1),\ i = 1, \ldots, n, &\qquad \boldsymbol{\theta}_1 \sim \pi_1(\boldsymbol{\theta}_1), \\ M_2: &\qquad Y_i \sim f_2(y_i \mid \boldsymbol{x}_{2,i}, \boldsymbol{\theta}_2),\ i = 1, \ldots, n, &\qquad \boldsymbol{\theta}_2 \sim \pi_1(\boldsymbol{\theta}_2), \end{aligned}$$

where, for example, the distributional assumption can coincide, $f_1(\cdot \mid \cdot) = f_2(\cdot \mid \cdot)$, and the model can differ only through the covariates which are including.

# Model assessment and comparison

A priori, we assume that $P(M_1) = \tau_1$ and $P(M_2) = \tau_2 = 1 - \tau_1$. The posterior probabilities of the two models can be obtained by first evaluating the **marginal density** for the data under $M_1$ and $M_2$, with

$$m_j(y_{1:n} \mid \boldsymbol{x}_{j,1:n}) = \int_\Theta \prod_{i=1}^n f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\theta}_j)\pi_j(\boldsymbol{\theta}_j)\mathrm{d}\boldsymbol{\theta}_j, \qquad j = 1, 2.$$

→ Recall that the previous measure how likely are $y_1, \ldots, y_n$ under model $M_j$.

Then, we can compute the **posterior model probabilities** in force of the Bayes' theorem

$$P(M_j \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \frac{m(y_{1:n} \mid \boldsymbol{x}_{j,1:n}, M_j)P(M_j)}{m(y_{1:n} \mid \boldsymbol{x}_{j,1:n})} = \frac{\tau_j m_j(y_{1:n} \mid \boldsymbol{x}_{j,1:n})}{\tau_1 m_1(y_{1:n} \mid \boldsymbol{x}_{1,1:n}) + \tau_2 m_2(y_{1:n} \mid \boldsymbol{x}_{2,1:n})},$$

for $j = 1, 2$, with $P(M_2 \mid y_{1:n}, \boldsymbol{x}_{2,1:n}) = 1 - P(M_1 \mid y_{1:n}, \boldsymbol{x}_{1,1:n})$.

We can use the previous probabilities, combined with the prior ones, to test if the two model are significant different from each other, by computing the Bayes factor

$$BF_{12} = \frac{\frac{P(M_1 \mid y_{1:n})}{P(M_2 \mid y_{1:n})}}{\frac{P(M_1)}{P(M_2)}} = \frac{m_1(y_1, \ldots, y_n)}{m_2(y_1, \ldots, y_n)}.$$

# Model assessment and comparison

But what if we have **more than two models**? A possible strategy is to consider jointly the posterior probabilities of different models. Suppose we have $k \geq 2$ models $M_1, \ldots, M_k$.

A priori, without any further information, we set the models **equally probable**

$$P(M_j) = \frac{1}{k}, \qquad j = 1, \ldots, k.$$

Similarly to before, we have different distributional assumptions $\{f_j(y_i \mid \boldsymbol{x}_{j,i}, \boldsymbol{\beta}_j)\}_{j=1}^k$ and prior assumptions $\pi_j(\boldsymbol{\beta}_j)$ that lead to different marginal distributions $\{m_j(y_i \mid \boldsymbol{x}_{j,1:n})\}_{j=1}^k$.

Hence, we can compute the **posterior probability** of each model as

$$P(M_j \mid y_{1:n}, \boldsymbol{x}_{j,1:n}) = \frac{m_j(y_{1:n} \mid \boldsymbol{x}_{j,1:n}) P(M_j)}{m(y_{1:n} \mid \boldsymbol{x}_{1:n})},$$

where $m(y_{1:n} \mid \boldsymbol{x}_{1:n}) = \sum_{j=1}^k m_j(y_{1:n} \mid \boldsymbol{x}_{j,1:n}) P(M_j)$.

One can then chose, for example, the model which has the largest posterior probability $P(M_j \mid y_{1:n}, \boldsymbol{x}_{j,1:n})$.

# Model assessment and comparison

## Example

```
set.seed(123); betatrue <- c(-4, 2, -4, 0)
z1 <- round(rnorm(100, 0, 1), digits = 1)
z2 <- round(rnorm(100, 0, 1), digits = 1)
z3 <- round(rnorm(100, 0, 1), digits = 1)
X <- cbind(rep(1, 100), z1, z2, z3)
tempprobs <- exp(as.vector(X %*% betatrue)) /
              (1 + exp(as.vector(X %*% betatrue)))
y <- sapply(tempprobs[,1], function(x) rbinom(1,1,x))
```

Consider three distinct logistic models, specifically

$$M_1 : y_i \sim Be(\theta_i), \qquad \theta_i = \mathrm{logit}(\beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2} + \beta_4 z_{i,3}), \qquad i = 1, \dots, n,$$
$$M_2 : y_i \sim Be(\theta_i), \qquad \theta_i = \mathrm{logit}(\beta_1 + \beta_2 z_{i,1} + \beta_3 z_{i,2}), \qquad i = 1, \dots, n,$$
$$M_3 : y_i \sim Be(\theta_i + \beta_2 z_{i,1}), \qquad \theta_i = \mathrm{logit}(\beta_1), \qquad i = 1, \dots, n.$$

- Write the augmented Gibbs sampler for the logistic model, which is also returning the pmf of each observation.
- Select the model which maximize the posterior probability.
- Test the selected model versus the simpler one.
- Repeat the previous points in STAN.

# Covariates selection

# Covariates selection

Suppose we now **fix** a certain **distribution** for the data, e.g. a specific model form. Further, we have a total of $p$ predictors available, with $\boldsymbol{x}_i^\mathsf{T} = (x_{i1}, \dots, x_{ip})$.

A reasonable question that we could ask to ourselves is weather we could find the **best model** for the response variable, given all the possible combination of covariates we can make.

Ideally, if we can estimate **all possible models** combining the covariates, we can then select the best one resorting to predictive information criteria as $\mathrm{WAIC}$ and $\mathrm{LPML}$.

- → Each covariate can be **included or not** in the model. We then have $2^p$ possible distinct models, given $p$ covariates.
- → When $p$ is large, it is **unfeasible** to estimate **all possible models** and then select the best one.

We need an alternative strategy to select the best subset of covariates, without estimating all possible models.

- → Many different regularization methods that also select covariates have been studied in the last decades.
- → In the following slides we consider an approach that has proven to be effective with many model strategies.

# Covariates selection

Here, we present the so called **spike-and-slab** approach.

- Usually, the marginal distribution of each regression coefficient is diffused on a real space, hence for the generic $j$th coefficient $\beta_j$, we have

$$P(\beta_j = 0) = 0.$$

- We **augment the prior distribution** of the regression coefficients defining a new prior specification which set **positive probability** on the 0 value (or on a neighborhood of 0).
- Such a prior is composed by two components
  - → **a spike one**, concentrating the mass;
  - → **a slab one**, which mimic the diffuseness of usual prior specifications.

In a general setting, we consider a model for which we have

$$g(\mathrm{E}[Y \mid \boldsymbol{x}]) = \boldsymbol{x}^\intercal \boldsymbol{\beta}, \qquad \boldsymbol{\beta} \sim \pi(\boldsymbol{\beta}).$$

→ Examples are the ordinary linear regression model, various GLMs, but the prior specification we are considering here can be embedded also in a mixed model environment.

→ $\pi(\boldsymbol{\beta})$ usually is assumed to be Gaussian, but other distributions can be considered here (e.g. the Laplace distribution of slide block 2).

# Covariates selection

At first, we consider an hierarchical specification as follows. We introduce a set of suitable augmented variables $\boldsymbol{\gamma}^{\intercal} = (\gamma_1, \dots, \gamma_p)$, where each $\gamma_j \in \{0, 1\}$, $j = 1, \dots, p$.

- The **augmented prior** can be written in a general form as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma})\pi(\boldsymbol{\gamma}).$$

- Each $\gamma_j$ describes the **inclusion or exclusion** of the $j$th covariate in the model, with

$$\gamma_j = \begin{cases} 1 \quad : & \text{if the } j\text{th covariate is included in the model,} \\ 0 \quad : & \text{otherwise.} \end{cases}$$

- Each one of the possible $2^p$ models is **uniquely identified** by a specific realization binary sequence $\boldsymbol{\gamma}^{\intercal} = (\gamma_1, \dots, \gamma_p)$.
- The prior is completed by setting $\gamma_j \sim Be(\theta_j)$, $j = 1, \dots, p$, $\theta_j \sim \pi(\theta_j)$, where the latter is not mandatory but improve the model flexibility, and $\theta_j$ denotes the probability that $\boldsymbol{\beta}_j$ is large enough to be included in the model.

We have a problem here, the dimension of the parameter space change as far as we are including or excluding covariates, as well for the prior dimension.

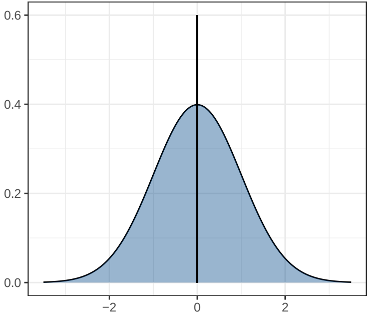
→ There exist suitable computational techniques to deal with this problem.

→ However, we can consider a similar prior specification, but more tractable.

# Covariates selection

Hence, instead of the augmented prior of before, we use the augmented variables $\gamma_j$s to construct a **mixture prior specification** for each regression coefficient.

- We want to have a diffuse prior apart from 0, following a specific distribution.
- We want to a positive probability for the coefficient of being equal (or close) to 0.



The prior specification we consider is then

$$\begin{aligned}\beta_j \mid \gamma_j &\overset{\text{ind}}{\sim} (1-\gamma_j)\delta_0 + \gamma_j N(0, \tau_j^2),\\ \gamma_j \mid \theta_j &\overset{\text{ind}}{\sim} Be(\theta_j),\\ \theta_j &\overset{\text{ind}}{\sim} \pi(\theta_j),\end{aligned}$$

for $j = 1, \ldots, p$, where $\delta_0$ denotes a Dirac measure in 0.

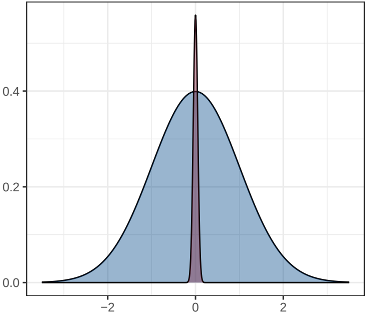
- Such a prior is flexible to cover the $2^p$ possible models, while being tractable.
- If $\theta_j = 0.5$ for all $j = 1, \ldots, p$, we are setting an uniform prior over all the possible $2^p$ models.

- However, we cannot use discrete prior distribution (such as the Dirac measure) in STAN...

# Covariates selection

We can consider a slightly relaxed version of the previous prior, by combining together **two Gaussian distributions**.

- A Gaussian distribution with large variance, which models actually coefficients different from 0.
- A Gaussian distribution with small variance, which set mass on a neighborhood of 0.



The prior specification is the following

$$\beta_j \mid \gamma_j \overset{\text{ind}}{\sim} (1-\gamma_j)N(0, c_j\tau_j^2) + \gamma_j N(0, \tau_j^2),$$

$$\gamma_j \mid \theta_j \overset{\text{ind}}{\sim} Be(\theta_j),$$

$$\theta_j \overset{\text{ind}}{\sim} \pi(\theta_j),$$

for $j = 1, \ldots, p$, where $\delta_0$ denotes a Dirac measure in 0 and $c_j > 0$ is small enough.

- The mixture prior distribution for $\beta_j$ is now manageable in STAN.
- Once we collect a sample from the posterior distribution, we can check how many times each regression coefficient has been sampled from the spike or the slab component.
- This approach is also called **stochastic search variable selection**.

# Covariates selection

A key quantity is the set of values $(-\kappa_j, \kappa_j)$ where **the spike components dominates the slab components**. The bounds of this set are given by $k_j = \tau_j \epsilon_j$, where

$$\epsilon_j = \sqrt{2 \frac{\log(c_j) c_j^2}{c_j^2 - 1}}.$$

→ When we sample a value $\beta_j \in (-\kappa_k, \kappa_k)$, we say that is close enough for zero.

As already mentioned, the parameter $\boldsymbol{\gamma}^\intercal = (\gamma_1, \dots, \gamma_p)$ **uniquely identifies** a specific model. Intuitively, we can compare models by looking at the specific **posterior probabilities**

$$P(\boldsymbol{\gamma} \mid y_{1:n}, \boldsymbol{x}_{1:n}) = \frac{m(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma})}{\sum_{\gamma \in \{0,1\}^p} m(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\gamma}) \pi(\boldsymbol{\gamma})},$$

where

$$m(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\gamma}) = \int_{\mathbb{R}^p} \mathrm{L}(y_{1:n} \mid \boldsymbol{x}_{1:n}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \pi(\boldsymbol{\beta} \mid \boldsymbol{\gamma}) \mathrm{d}\boldsymbol{\beta}$$

denotes the marginal distribution of the data including the covariates in $\boldsymbol{\gamma}$.

# Covariates selection

Once we simulate a **sample from the posterior distribution** $\left(\boldsymbol{\beta}^{(r)}, \boldsymbol{\gamma}^{(r)}\right)_{r=1}^{R}$, we are interested into **select a specific set of covariates**, in force of the posterior empirical information we have. We can resort to different strategies, specifically the following.

- **Highest posterior probability (HPD)**, we choose the model with the highest posterior probability, i.e.

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma} \in \{0,1\}^p}{\arg\max}\, \pi(\boldsymbol{\gamma} \mid y_{1:n}, \boldsymbol{x}_{1:n}) \approx \underset{\boldsymbol{\gamma} \in \{0,1\}^p}{\arg\max}\, \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}_{[\boldsymbol{\gamma}^{(r)} = \boldsymbol{\gamma}]},$$

  i.e. the mode of the posterior distribution.

- **Median probability model (MPM)**, we choose the covariates having posterior probabilities of being included greater than 0.5, i.e.

$$\gamma_j \text{ such that } \pi(\gamma_j \mid y_{1:n}, \boldsymbol{x}_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}_{[\gamma_j = 1]} > 0.5.$$

- **Hard shrinkage (HS)**, we choose only the covariates that are always included in the model, i.e.

$$\gamma_j \text{ such that } \pi(\gamma_j \mid y_{1:n}, \boldsymbol{x}_{1:n}) \approx \frac{1}{R} \sum_{r=1}^{R} \mathbf{1}_{[\gamma_j = 1]} = 1.$$

# Covariates selection

## Example

Consider the following data generating process

```
set.seed(123); betatrue <- c(2, 0, 0, -3, 4, 0, -1, -2, rep(0, 12))
X <- cbind(rep(1, 100),
        round(matrix(rnorm(1900), ncol = 19), digits = 2))
tempmeans <- X %*% betatrue
y <- sapply(tempmeans, function(x) rnorm(1, x, 1))
```

Consider a generic linear model $y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2)$. We further set a priori

$$\sigma^2 \sim IG(2,2), \qquad \beta_j \sim (1-\theta_j)N(0, 0.001 \times 10^3) + \theta_j N(0, 10^3),\ j = 1, \dots, p,$$

with $\theta \sim Beta(1,1)$.

- Write the STAN code to sample from the posterior distribution of the model.
- Calculate the intersection point of the Gaussian distributions in the prior specification of $\beta_j$.
- Provide the HPD, MPM and HS estimates of the best subset of covariates for the model, and compare them with the covariates used to simulate the data.

# Relaxing the model specification

# Relaxing the model specification

Our model specification relies on **our prior assumption**, which is then updated in our posterior belief.

- It is possible that the prior guess we are taking it is too strongly concentrated on a specific part of the parameter space.
  - → In the worst scenario, it is strongly concentrated in the wrong part of the parameter space.
- In force of that, sometimes models are too sensible to the parameter specification.

A simple but effective way to avoid this problem is to include one extra hierarchical level in your model, by setting a prior distributions on the main hyperparameters of your model, ideally

$$
\begin{aligned}
Y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta} &\sim f(y_i \mid \boldsymbol{x}_i, \boldsymbol{\beta}), \\
\boldsymbol{\beta} \mid \boldsymbol{\theta} &\sim \pi(\boldsymbol{\beta} \mid \boldsymbol{\theta}), \\
\boldsymbol{\theta} &\sim \pi(\boldsymbol{\theta}).
\end{aligned}
$$

- Be careful to match the support of $\boldsymbol{\theta}$ with the support of the distribution assumption you are taking.
- In a few situations, you still preserve conjugacy.

# Relaxing the model specification

## Example

Consider the following data generating process

```
set.seed(123); betatrue <- c(2, -3, 4, -2)
X <- cbind(rep(1, 100),
        round(matrix(rnorm(300), ncol = 19), digits = 2))
tempmeans <- X %*% betatrue
y <- sapply(tempmeans, function(x) rnorm(1, x, 1))
```

Consider a generic linear model $y_i = \boldsymbol{x}_i^\intercal \boldsymbol{\beta} + \epsilon_i$, with $\epsilon_i \sim N(0, \sigma^2)$.

- Write the STAN code to sample from the posterior distribution of the model, considering a priori $\sigma^2 \sim IG(100, 1)$, $\quad \boldsymbol{\beta} \sim N(\mathbf{25}, \mathrm{diag}(0.1))$.
- Write now the STAN code to sample from the posterior distribution of the model, considering a priori

  $$\sigma^2 \sim IG(a_0, b_0), \qquad \boldsymbol{\beta} \sim N(\boldsymbol{b}_0, \Sigma_0),$$

  with hyperpriors

  $$a_0 \sim Gamma(0.1, 0.1), \qquad b_0 \sim Gamma(0.1, 0.1),$$
  $$\boldsymbol{b}_0 \sim N(\mathbf{0}, \mathrm{diag}(10^3)), \qquad \Sigma_0 \sim Wishart(6, \mathrm{diag}(1)).$$