Article:
Nayek, R. orcid.org/0000-0003-4277-8382, Fuentes, R., Worden, K. et al. (1 more author) (2021) On spike-and-slab priors for Bayesian equation discovery of nonlinear dynamical systems via sparse linear regression. Mechanical Systems and Signal Processing, 161. 107986. ISSN 0888-3270

https://doi.org/10.1016/j.ymssp.2021.107986

Supplementary Material

On spike-and-slab priors for Bayesian equation discovery of nonlinear dynamical systems via sparse linear regression

R. Nayek, R. Fuentes, K. Worden, E.J. Cross

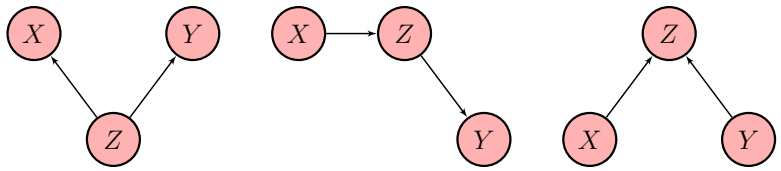
This document comprises the derivation of the conditional probability distributions in the Gibbs sampling scheme used for Discontinuous Spike-and-Slab (DSS) priors. The spike-and-slab prior has a hierarchical form that can be expressed in the form of a directed acyclic graph (DAG), and hence, it will be useful to understand the play of conditional independencies between the random variables in a DAG, before looking to derive the conditional probability distributions of the Gibbs sampler.

1. Conditional independence in DAGs

Given three random variables, say X, Y and Z,

$$X \perp Y \mid Z \tag{1}$$

implies that X and Y are probabilistically *independent* given $Z = z$. Conditional independencies between random variables are easily visualised using DAGs, and whether or not two random variables (depicted by the nodes in a DAG) are conditionally independent is decided based on their structure of their connection i.e. the directed edges of the DAG. Figure 1 shows three different types of DAG structures.



$X \perp Y \mid Z$ $\quad$ $X \perp Y \mid Z$ $\quad$ $X \not\perp Y \mid Z$

(a) (b) (c)

Figure 1: Three different types of directed acyclic graph structures.

In the first two DAG structures, 1(a) and 1(b), the conditioning on Z renders Y independent of X, and therefore, graph structures similar to cases (a) and (b) would imply $X \perp Y \mid Z$. For DAG structure 1(c), X and Y are marginally independent of each other, that is $X \perp Y$, however, when conditioned on Z, X and Y become dependent, i.e. $X \not\perp Y \mid Z$. The notion of conditional independence in DAGs would greatly aid in finding the conditional probability distributions feeding the Gibbs sampler.

2. Derivation of the conditional probability distributions of Gibbs sampler for DSS priors

2.1. DSS prior model specification

The DSS prior model used for linear regression with dictionary $\mathbf{D} \in \mathbb{R}^{N \times P}$ is summarised by the following set of relations:

$$p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) = \mathcal{N}\left(\mathbf{D}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_N\right) \tag{2a}$$

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{z}\right) = p_{\text{slab}}\left(\boldsymbol{\theta}_r\right) \prod_{i:z_i=0} p_{\text{spike}}\left(\theta_i\right) \tag{2b}$$

$$p_{\text{spike}}\left(\theta_i\right) = \delta_0 \tag{2c}$$

$$p_{\text{slab}}\left(\boldsymbol{\theta}_r \mid \sigma^2, v_s\right) = \mathcal{N}\left(\mathbf{0}, \sigma^2 v_s \mathbf{A}_{0,r}\right) \tag{2d}$$

$$p\left(v_s\right) = \mathcal{IG}\left(a_v, b_v\right) \tag{2e}$$

$$p\left(z_i \mid p_0\right) = \text{Bern}(p_0) \tag{2f}$$

$$p\left(p_0\right) = \text{Beta}(a_p, b_p) \tag{2g}$$

$$p\left(\sigma^2\right) = \mathcal{IG}\left(a_\sigma, b_\sigma\right) \tag{2h}$$

Here, $\boldsymbol{\theta}_r \in \mathbb{R}^{r \times 1}$ denotes the set of components of $\boldsymbol{\theta}$ for which the corresponding z_is take values 1. The relations between the random variables can be visualised as in Figure 2.

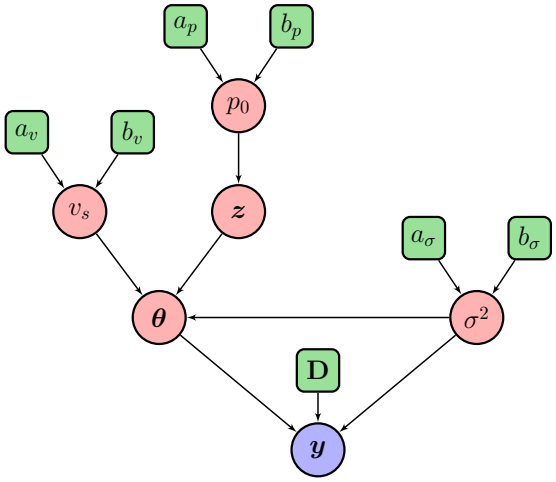


Figure 2: Graph of the hierarchical spike-and-slab model for linear regression; the variables in circles represent random variables, while those in squares represent deterministic parameters.

Due to the DAG structure, the joint probability distribution over all random variables factorises as,

$$p\left(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, p_0, v_s, \sigma^2\right) = p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) p\left(\boldsymbol{z} \mid p_0\right) p\left(p_0\right) p\left(v_s\right) p\left(\sigma^2\right) \tag{3}$$

Note the conditioning on deterministic hyperparameters $a_v, b_v, a_p, b_p, a_\sigma, b_\sigma$ and the dictionary $\mathbf{D}$ has been suppressed. The Gibbs sampler needs closed-form expressions of the conditional probability distributions of each random variable given the data and all other random variables. However, due to the DAG structure, the conditional probability distributions of the random variables become conditionally independent of certain variables:

$$p\left(p_0 \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, v_s, \sigma^2\right) = p\left(p_0 \mid \boldsymbol{z}\right) \tag{4a}$$

$$p\left(v_s \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, p_0, \sigma^2\right) = p\left(v_s \mid \boldsymbol{\theta}, \boldsymbol{z}, \sigma^2\right) \tag{4b}$$

$$p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, p_0, v_s\right) = p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, v_s\right) \tag{4c}$$

$$p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}, p_0, v_s, \sigma^2\right) = p\left(\boldsymbol{\theta} \mid \boldsymbol{y}, \boldsymbol{z}, v_s, \sigma^2\right) \tag{4d}$$

$$p\left(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\theta}, p_0, v_s, \sigma^2\right) = p\left(\boldsymbol{z} \mid \boldsymbol{\theta}, p_0, v_s, \sigma^2\right) \tag{4e}$$

At this point, it must be mentioned that the Dirac-delta spike distribution in DSS priors creates a nuisance in the Gibbs sampling scheme; it prevents the creation of an irreducible Markov chain required for the chain to reach a stationary distribution. Specifically, when $z_i = 0$, the new value for θ_i will be $\theta_i = 0$, which will

in turn result in $z_i = 0$ in a new draw, and the process keeps repeating itself. In other words, the chain has absorbing states. This problem can be avoided by *marginalising* or integrating over $\boldsymbol{\theta}$ to get rid of the Dirac-delta function. Therefore, in the derivations of the conditional probability distributions for Eqs. 4c and 4e, the procedure of marginalising over $\boldsymbol{\theta}$ is performed to obtain an irreducible chain. As a result of marginalisation over $\boldsymbol{\theta}$, the conditional probabilities Eqs. 4c and 4e become dependent upon $\boldsymbol{y}$, and their relations can be re-expressed as,

$$p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, p_0, v_s\right) = p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{z}, v_s\right) \tag{5a}$$

$$p\left(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{\theta}, p_0, v_s, \sigma^2\right) = p\left(\boldsymbol{z} \mid \boldsymbol{y}, p_0, v_s, \sigma^2\right) \tag{5b}$$

2.2. Derivations of conditional probability distributions

In this section, the conditional sampling distributions used in Gibbs sampling with DSS priors are derived.

2.2.1. Sampling distribution of p_0

$$\begin{aligned}
p\left(p_0 \mid \boldsymbol{z}\right) &\propto p\left(\boldsymbol{z} \mid p_0\right) p\left(p_0\right) \\
&\propto \left(\prod_{i=1}^{P} p_0^{z_i} \left(1-p_0\right)^{1-z_i}\right) p_0^{a_p-1} \left(1-p_0\right)^{b_p-1} \qquad \left[\because z_i\text{'s are independent}\right] \\
&\propto p_0^{a_p+s_z} \left(1-p_0\right)^{b_p+P-s_z} \qquad \left[\text{where } s_z = \sum_{i=1}^{P} z_i\right]
\end{aligned}$$

$$\boxed{p_0 \mid \boldsymbol{z} \sim \text{Beta}\left(a_p + s_z,\ b_p + P - s_z\right)} \tag{6}$$

2.2.2. Sampling distribution of v_s

$$\begin{aligned}
p\left(v_s \mid \boldsymbol{\theta}, \boldsymbol{z}, \sigma^2\right) &\propto p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) p\left(v_s\right) p\left(\boldsymbol{z}\right) p\left(\sigma^2\right) \\
&\propto p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) p\left(v_s\right) \qquad \left[\because p\left(\boldsymbol{z}\right) \text{ and } p\left(\sigma^2\right) \text{ are constants w.r.t. } v_s\right] \\
&\propto \mathcal{N}\left(\boldsymbol{\theta}_r \mid \boldsymbol{0}, \sigma^2 v_s \mathbf{A}_{0,r}\right)\ \mathcal{IG}\left(a_v,\ b_v\right) \\
&\propto \frac{1}{\left(v_s\sigma^2\right)^{r/2} \left|\mathbf{A}_{0,r}\right|^{1/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T \mathbf{A}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2 v_s}\right)\ v_s^{-a_v-1} \exp\left(-\frac{b_v}{v_s}\right) \\
&\propto v_s^{-\left(a_v+\frac{r}{2}\right)-1} \exp\left(-\frac{b_v + \frac{\boldsymbol{\theta}_r^T \mathbf{A}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2}}{v_s}\right)
\end{aligned}$$

$$\boxed{v_s \mid \boldsymbol{\theta}, \boldsymbol{z}, \sigma^2 \sim \mathcal{IG}\left(a_v + \frac{r}{2},\ b_v + \frac{\boldsymbol{\theta}_r^T \mathbf{A}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2}\right)} \tag{7}$$

Note, r is the number of components of $\boldsymbol{\theta}$ that fall in the slab and is equal to the number of non-zero components of $\boldsymbol{z}$. Since z_is are binary variables, $r = s_z = \sum_{i=1}^{P} z_i$. When all z_is are zero, v_s is sampled from the prior $\mathcal{IG}\left(a_v, b_v\right)$.

2.2.3. Sampling distribution of σ^2

The derivation of conditional distribution of σ^2 will involve integrating out the parameter $\boldsymbol{\theta}$ to avoid dealing with the Dirac-delta spike distribution (as mentioned before).

$$\begin{aligned}
p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{z}, v_s\right) &\propto \int p\left(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta} \\
&\propto \left(\int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta}\right) p\left(\boldsymbol{z}\right) p\left(v_s\right) p\left(\sigma^2\right) \\
&\propto \left(\int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta}\right) p\left(\sigma^2\right) \qquad \left[\because p\left(\boldsymbol{z}\right) \text{ and } p\left(v_s\right) \text{ are constants w.r.t. } \sigma^2\right]
\end{aligned}$$

Upon integration, the components of $\boldsymbol{\theta}$ belonging to the spike distribution are evaluated at zero and the rest of the components of $\boldsymbol{\theta}$ that belong to the slab (i.e. $\boldsymbol{\theta}_r$) remains to be integrated. Expanding the integrand $p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right)$, with $\boldsymbol{\theta}$ now replaced by $\boldsymbol{\theta}_r$, one obtains,

$$\frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{y}-\mathbf{D}_r\boldsymbol{\theta}_r\right)^T\left(\boldsymbol{y}-\mathbf{D}_r\boldsymbol{\theta}_r\right)}{2\sigma^2}\right) \quad \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}}{\left(2\pi v_s\sigma^2\right)^{r/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T\mathbf{A}_{0,r}^{-1}\boldsymbol{\theta}_r}{2\sigma^2 v_s}\right)$$

Simplifying and rearranging the terms involving $\boldsymbol{\theta}_r$ leads to,

$$\begin{aligned}
&\implies \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}}{\left(2\pi v_s\sigma^2\right)^{r/2}} \exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}+\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_r-2\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2\sigma^2}\right) \\
&\implies \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}}{\left(2\pi v_s\sigma^2\right)^{r/2}} \exp\left(-\frac{\left(\boldsymbol{\theta}_r-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}_r-\boldsymbol{\mu}\right)}{2\sigma^2}\right) \exp\left(-\frac{\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}{2\sigma^2}\right)
\end{aligned}$$

where $\boldsymbol{\Sigma}^{-1} = \left(\mathbf{D}_r^T\mathbf{D}_r + v_s^{-1}\mathbf{A}_{0,r}^{-1}\right)$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{D}_r^T\boldsymbol{y}$. On integrating out $\boldsymbol{\theta}_r$ from the above expression, the conditional distribution of σ^2 reduces to:

$$\begin{aligned}
p\left(\sigma^2 \mid \boldsymbol{y}, \boldsymbol{z}, v_s\right) &\propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}{2\sigma^2}\right) \quad p\left(\sigma^2\right) \\
&\propto \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}{2\sigma^2}\right) \left(\sigma^2\right)^{-a_\sigma-1} \exp\left(-\frac{b_\sigma}{\sigma^2}\right) \\
&\propto \left(\sigma^2\right)^{-\left(a_\sigma+\frac{N}{2}\right)-1} \exp\left(-\frac{b_\sigma+\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)}{\sigma^2}\right)
\end{aligned}$$

$$\boxed{\sigma^2 \mid \boldsymbol{y}, \boldsymbol{z}, v_s \sim \mathcal{IG}\left(a_\sigma + \frac{N}{2},\ b_\sigma + \frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)\right)} \tag{8}$$

Note that when all z_is are equal to zero, σ^2 is sampled from $\mathcal{IG}\left(a_\sigma + \frac{N}{2},\ b_\sigma + \frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}\right)\right)$.

2.2.4. Sampling distribution of $\boldsymbol{\theta}$

The components of $\boldsymbol{\theta}$ that correspond to $z_i = 0$ are set to zero (as they belong to the spike Dirac-delta distribution). The conditional distribution for the rest of the components belonging to the slab, represented by $\boldsymbol{\theta}_r$, can be derived as follows:

$$\begin{aligned}
p\left(\boldsymbol{\theta}_r \mid \boldsymbol{y}, v_s, \sigma^2\right) &\propto \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{D}_r\boldsymbol{\theta}_r, \sigma^2\mathbf{I}_N\right) \ \mathcal{N}\left(\boldsymbol{\theta}_r \mid \mathbf{0}, \sigma^2 v_s\mathbf{A}_{0,r}\right) \\
&\propto \exp\left(-\frac{\left(\boldsymbol{y}-\mathbf{D}_r\boldsymbol{\theta}_r\right)^T\left(\boldsymbol{y}-\mathbf{D}_r\boldsymbol{\theta}_r\right)}{2\sigma^2}\right) \ \exp\left(-\frac{\boldsymbol{\theta}_r^T\mathbf{A}_{0,r}^{-1}\boldsymbol{\theta}_r}{2\sigma^2 v_s}\right) \\
&\propto \exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}+\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\theta}_r-2\boldsymbol{\theta}_r^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{2\sigma^2}\right) \\
&\propto \exp\left(-\frac{\left(\boldsymbol{\theta}_r-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\theta}_r-\boldsymbol{\mu}\right)}{2\sigma^2}\right)
\end{aligned}$$

where $\boldsymbol{\Sigma}^{-1} = \left(\mathbf{D}_r^T\mathbf{D}_r + v_s^{-1}\mathbf{A}_{0,r}^{-1}\right)$ and $\boldsymbol{\mu} = \boldsymbol{\Sigma}\mathbf{D}_r^T\boldsymbol{y}$.

$$\boxed{\boldsymbol{\theta}_r \mid \boldsymbol{y}, v_s, \sigma^2 \sim \mathcal{N}\left(\boldsymbol{\mu},\ \sigma^2\boldsymbol{\Sigma}\right)} \tag{9}$$

Note that the explicit conditional dependence of $\boldsymbol{\theta}_r$ on $\boldsymbol{z}$ has been suppressed, because it is understood that $\boldsymbol{\theta}_r$ corresponds to those components of $\boldsymbol{\theta}$ for which the corresponding z_is take values of unity.

2.2.5. Sampling distribution of $\boldsymbol{z}$

In the derivation of sampling distribution of $\boldsymbol{z}$, the parameter $\boldsymbol{\theta}$ is marginalised to avoid dealing with the Dirac-delta functions; as a consequence, the conditional probability distributions of z_is become dependent on $\boldsymbol{y}$. Additionally, the sampling distribution is marginalised over the noise-variance parameter σ^2 as this step leads to a quicker convergence of the Markov chains. As such, the resulting sampling distribution of $\boldsymbol{z}$, $p\left(\boldsymbol{z} \mid \boldsymbol{y}, p_0, v_s\right)$, is arrived at after performing integrations over both $\boldsymbol{\theta}$ and σ^2:

$$p\left(\boldsymbol{z} \mid \boldsymbol{y}, p_0, v_s\right) \propto \left(\int \left(\int p\left(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta} \right) p\left(\sigma^2\right) d\sigma^2 \right) p\left(\boldsymbol{z} \mid p_0\right) \tag{10}$$

Since individual components of $\boldsymbol{z}$ are independent of each other, the conditional probability distributions of z_i are derived separately. The computation of conditional probability distribution over z_i requires computing the probability of $z_i = 1$ compared to $z_i = 0$, given the same values of $\boldsymbol{z}_{-i}$, v_s and p_0; the term $\boldsymbol{z}_{-i}$ denotes $\boldsymbol{z}$ with it's i^{th} component removed. Denote by ξ_i the probability with which one samples $z_i = 1$, then ξ_i can be expressed as:

$$\begin{aligned}
\xi_i &= \frac{p\left(z_i = 1 \mid \boldsymbol{y}, \boldsymbol{z}_{-i}, v_s, p_0\right)}{p\left(z_i = 1 \mid \boldsymbol{y}, \boldsymbol{z}_{-i}, v_s, p_0\right) + p\left(z_i = 0 \mid \boldsymbol{y}, \boldsymbol{z}_{-i}, v_s, p_0\right)} \\
&= \frac{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right) p\left(z_i = 1 \mid p_0\right)}{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right) p\left(z_i = 1 \mid p_0\right) + p\left(\boldsymbol{y} \mid z_i = 0, \boldsymbol{z}_{-i}, v_s\right) p\left(z_i = 0 \mid p_0\right)} \\
&= \frac{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right) p_0}{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right) p_0 + p\left(\boldsymbol{y} \mid z_i = 0, \boldsymbol{z}_{-i}, v_s\right) \left(1 - p_0\right)} \\
&= \frac{p_0}{p_0 + \frac{p\left(\boldsymbol{y} \mid z_i = 0, \boldsymbol{z}_{-i}, v_s\right)}{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right)} \left(1 - p_0\right)} = \frac{p_0}{p_0 + R_i \left(1 - p_0\right)} \qquad \left[\text{where } R_i = \frac{p\left(\boldsymbol{y} \mid z_i = 0, \boldsymbol{z}_{-i}, v_s\right)}{p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right)}\right]
\end{aligned} \tag{11}$$

As seen above, the evaluation of ξ_i requires the computation of the marginal likelihood $p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s\right)$, which will be derived next. The derivation of the marginal likelihood involves integrating the likelihood with respect to the distribution over $\boldsymbol{\theta}$, followed by that over σ^2.

Marginalisation over $\boldsymbol{\theta}$

$$\begin{aligned}
& p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s, \sigma^2\right) \\
&= \int p\left(\boldsymbol{y}, \boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta} \\
&= \int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}, \sigma^2\right) p\left(\boldsymbol{\theta} \mid \boldsymbol{z}, v_s, \sigma^2\right) d\boldsymbol{\theta} \\
&= \int p\left(\boldsymbol{y} \mid \boldsymbol{\theta}_r, \sigma^2\right) p\left(\boldsymbol{\theta}_r \mid v_s, \sigma^2\right) d\boldsymbol{\theta}_r \qquad [\because \text{Dirac-delta functions are integrated out}] \\
&= \int \mathcal{N}\left(\boldsymbol{y} \mid \mathbf{D}_r \boldsymbol{\theta}_r, \sigma^2 \mathbf{I}_N\right) \ \mathcal{N}\left(\boldsymbol{\theta}_r \mid \mathbf{0}, \sigma^2 v_s \mathbf{A}_{0,r}\right) d\boldsymbol{\theta}_r \\
&= \int \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{y} - \mathbf{D}_r \boldsymbol{\theta}_r\right)^T \left(\boldsymbol{y} - \mathbf{D}_r \boldsymbol{\theta}_r\right)}{2\sigma^2} \right) \ \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}}{\left(2\pi v_s \sigma^2\right)^{r/2}} \exp\left(-\frac{\boldsymbol{\theta}_r^T \mathbf{A}_{0,r}^{-1} \boldsymbol{\theta}_r}{2\sigma^2 v_s} \right) d\boldsymbol{\theta}_r \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}}{\left(2\pi v_s \sigma^2\right)^{r/2}} \exp\left(-\frac{\left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{a}_N^T \mathbf{A}_N^{-1} \boldsymbol{a}_N\right)}{2\sigma^2} \right) \int \exp\left(-\frac{\left(\boldsymbol{\theta}_r - \boldsymbol{a}_N\right)^T \mathbf{A}_N^{-1} \left(\boldsymbol{\theta}_r - \boldsymbol{a}_N\right)}{2\sigma^2} \right) d\boldsymbol{\theta}_r \\
&= \frac{1}{\left(2\pi\sigma^2\right)^{N/2}} \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2} \left(\det\left(\mathbf{A}_N\right)\right)^{1/2}}{\left(v_s\right)^{r/2}} \exp\left(-\frac{\left(\boldsymbol{y}^T \boldsymbol{y} - \boldsymbol{a}_N^T \mathbf{A}_N^{-1} \boldsymbol{a}_N\right)}{2\sigma^2} \right)
\end{aligned}$$

where $\mathbf{A}_N = \left(\mathbf{D}_r^T \mathbf{D}_r + v_s^{-1} \mathbf{A}_{0,r}^{-1}\right)^{-1}$ and $\boldsymbol{a}_N = \mathbf{A}_N \mathbf{D}_r^T \boldsymbol{y}$.

Marginalisation over σ^2

$$
\begin{aligned}
& p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s\right) \\
=& \int p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s, \sigma^2\right) p\left(\sigma^2\right) d\sigma^2 \\
=& \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}\left(\det\left(\mathbf{A}_N\right)\right)^{1/2}}{(2\pi)^{N/2}\left(v_s\right)^{r/2}} \int \frac{1}{\left(\sigma^2\right)^{N/2}} \exp\left(-\frac{\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{a}_N^T\mathbf{A}_N^{-1}\boldsymbol{a}_N\right)}{2\sigma^2}\right) \; \mathcal{IG}\left(a_\sigma, b_\sigma\right) \; d\sigma^2 \\
=& \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}\left(\det\left(\mathbf{A}_N\right)\right)^{1/2}}{(2\pi)^{N/2}\left(v_s\right)^{r/2}} \frac{\left(b_\sigma\right)^{a_\sigma}}{\Gamma\left(a_\sigma\right)} \int \frac{1}{\left(\sigma^2\right)^{a_\sigma+N/2+1}} \exp\left(-\frac{b_\sigma+\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{a}_N^T\mathbf{A}_N^{-1}\boldsymbol{a}_N\right)}{\sigma^2}\right) d\sigma^2 \\
=& \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}\left(\det\left(\mathbf{A}_N\right)\right)^{1/2}}{(2\pi)^{N/2}\left(v_s\right)^{r/2}} \frac{\left(b_\sigma\right)^{a_\sigma}}{\Gamma\left(a_\sigma\right)} \frac{\Gamma\left(a_\sigma+\frac{N}{2}\right)}{\left(b_\sigma+\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{a}_N^T\mathbf{A}_N^{-1}\boldsymbol{a}_N\right)\right)^{\left(a_\sigma+\frac{N}{2}\right)}}
\end{aligned}
$$

where $\Gamma(\cdot)$ denotes the Gamma function. As such,

$$
p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s\right) = \begin{cases} \frac{\Gamma\left(a_\sigma+\frac{N}{2}\right)}{(2\pi)^{N/2}} \frac{\left(b_\sigma\right)^{a_\sigma}}{\Gamma\left(a_\sigma\right)} \frac{1}{\left(b_\sigma+\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}\right)\right)^{\left(a_\sigma+\frac{N}{2}\right)}} & \text{when all } z_i\text{s are zero} \\ \frac{\Gamma\left(a_\sigma+\frac{N}{2}\right)}{(2\pi)^{N/2}\left(v_s\right)^{r/2}} \frac{\left(b_\sigma\right)^{a_\sigma}}{\Gamma\left(a_\sigma\right)} \frac{\left(\det\left(\mathbf{A}_{0,r}^{-1}\right)\right)^{1/2}\left(\det\left(\mathbf{A}_N\right)\right)^{1/2}}{\left(b_\sigma+\frac{1}{2}\left(\boldsymbol{y}^T\boldsymbol{y}-\boldsymbol{a}_N^T\mathbf{A}_N^{-1}\boldsymbol{a}_N\right)\right)^{\left(a_\sigma+\frac{N}{2}\right)}} & \text{otherwise} \end{cases} \tag{12}
$$

Finally, the sampling distributions of z_i, $i = 1, \ldots, P$, can be computed using a Bernoulli distribution:

$$
\boxed{z_i \mid \boldsymbol{y}, v_s, p_0 \sim \text{Bern}\left(\xi_i\right)} \tag{13}
$$

with $\xi_i = \frac{p_0}{p_0 + R_i(1-p_0)}$ and $R_i = \frac{p(\boldsymbol{y}|z_i=0, \boldsymbol{z}_{-i}, v_s)}{p(\boldsymbol{y}|z_i=1, \boldsymbol{z}_{-i}, v_s)}$ calculated using Eq. (12). Lastly, it should be mentioned that to avoid numerical overflow errors, it is practical to use the logarithm of the marginal likelihoods $p\left(\boldsymbol{y} \mid \boldsymbol{z}, v_s\right)$ to compute $R_i = \exp\left\{\log\left[p\left(\boldsymbol{y} \mid z_i = 0, \boldsymbol{z}_{-i}, v_s\right)\right] - \log\left[p\left(\boldsymbol{y} \mid z_i = 1, \boldsymbol{z}_{-i}, v_s\right)\right]\right\}$.