Supplement to "Bayes Factor Asymptotics for Variable Selection in the Gaussian Process Framework"

Minerva Mukhopadhyay Sourabh Bhattacharya

This document is an addendum to the theory developed in the main manuscript (MB). This supplementary material is organized as follows.

In Section S-1, we numerically validate the results of MB, in the contexts of linear regression (LR), Gaussian process regression (GPR) and AR(1) process regression (AR-1).

Next we consider the problem of Bayes factor based variable selection from among $2^p - 1$ available models. In this regard, in Section S-2 we introduce our TTMCMC sampler for general Bayesian variable selection problems. The method of computation of Bayes factors using TTMCMC samples is detailed in Section S-3. In Section S-4 we provide the proof of convergence of our TTMCMC sampler.

In Section S-5 we provide the details of our TTMCMC based variable selection experiments in the contexts of LR, GPR, and AR-1.

In Section S-6 we address variable selection among a set of 4088 covariates in a real, riboflavin dataset, using our Bayes factor oriented TTMCMC methodology, considering both linear and Gaussian process regression, and obtain interesting insights with respect to existing results on variable selection in the same dataset obtained using linear regression and classical methods.

Finally, in Section S-7, we provide the proof of the lemmas and results stated in the MB.

S-1 Direct validation of the theoretical results using simulation experiments

S-1.1 Linear regression

Here we assume $y_i = \boldsymbol{\beta}_{\mathbf{s}_0}^T \mathbf{x}_{i,\mathbf{s}_0} + \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0,1)$. As stated above the covariates are generated from scaled t_3 distribution, where scale matrix Σ_0 is AR(1) structured, with $\rho = 0.25$. We assign Zellner's g-prior on the regression coefficients $\boldsymbol{\beta}_{\mathbf{s}}$, with $\boldsymbol{\beta}_{0,\mathbf{s}} = (1/p, \ldots, 1/p)$, $\sigma^2_{\boldsymbol{\beta}} = 1$ and $g = 10$. As set of $|\mathbf{s}_0|$ covariates are chosen at random, and the values of the corresponding coefficients are chosen from an Uniform$(0,1)$ distribution.

The results are summarized in Figure S-1. Note that, the *supermodel* has exactly one extra variable and the *altered model* has exactly one variable different from the true model. Even for such small changes, the Bayes factor identifies the true model efficiently. Further, as the size of the true model increases, Bayes factor becomes more efficient.

S-1.2 Gaussian process with squared exponential kernel

Next we generate data from Gaussian process with squared exponential kernel as given in (11). We choose $D_{\mathbf{s}} = \text{diag}\{10, \ldots, 10\}$ for all $\mathbf{s}$, and $\sigma_f^2 = 1$. We choose a constant mean function, $\boldsymbol{\mu}_{\mathbf{s}} = \text{logistic}\,(\mathbf{x}'_{\mathbf{s}}\boldsymbol{\beta}_{\mathbf{s}})$ for all $\mathbf{s}$. Note that, the assumptions (A1)-(A3) are satisfied by these choices of the parameters. The coefficients, $\boldsymbol{\beta}_{\mathbf{s},j}$, are generated randomly from independent Uniform$(-0.5, 0.5)$ distributions.

As before the covariates are generated from scaled t_3 distribution, where scale matrix Σ_0 is AR(1) structured, with $\rho = 0.1$. In this case the *supermodel* has $k = 5$ more covariates, and the *altered model* has $k = 5$ different covariates than the true model. These covariates are randomly selected from the pool of $p - |\mathbf{s}_0|$ covariates. Figure S-2 shows the performance of the Bayes factor as n grows. Observe that, unlike both linear regression and AR(1) regression, the Bayes factor detects the true model much faster when some covariates are altered, than a supermodel.

S-1.3 Autoregressive model

The response is now generated from AR(1) model (19) with $\rho = 0.25$. The distribution of the covariates, choice of prior on $\boldsymbol{\beta}_{\mathbf{s}}$, and the definition of *supermodel* and *altered model* are same as that in Section S-1.1.

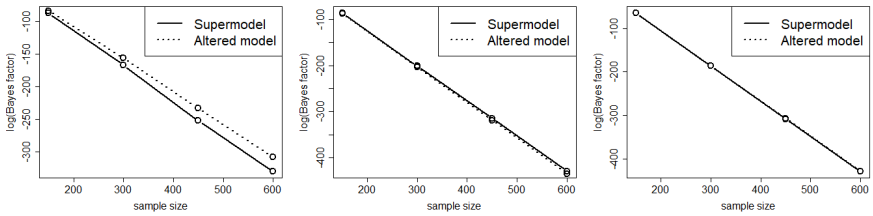


Figure S-1: Line diagram showing how the Bayes factors decrease as sample sizes increase in linear regression when $|\mathbf{s}_0| = 10$ (left panel), $|\mathbf{s}_0| = 40$ (middle panel), and $|\mathbf{s}_0| = 40$ (right panel)

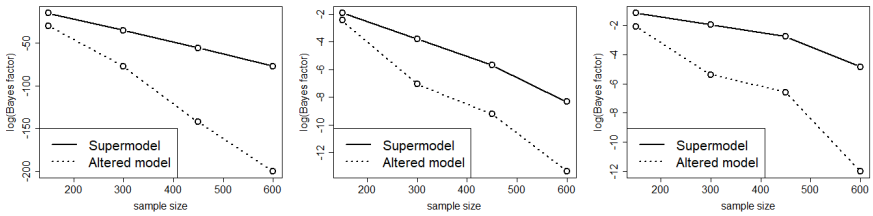


Figure S-2: Line diagram showing how the Bayes factors decrease as sample sizes increase in GP with squared exponential kernel when $|\mathbf{s}_0| = 10$ (left panel), $|\mathbf{s}_0| = 40$ (middle panel), and $|\mathbf{s}_0| = 40$ (right panel)

As the true value of ρ is not known, we numerically find the integrated marginal likelihood of the true model and the competing model, considering an Uniform$(-1, 1)$ prior on ρ. The integrated Bayes factor is the ratio of the integrated likelihood of the competing and the true model.

The results are summarized in Figure S-3. As in the case of linear model, Bayes factor efficiently captures the true model even when the competing model is the closest one to the truth.

S-1.4 Misspecified models

Now we compare two nested supermodels of the true model, $\mathcal{M}_{\mathbf{s}_1} \subset \mathcal{M}_{\mathbf{s}_2}$, with dimensions $|\mathbf{s}_1| = k_1$ and $|\mathbf{s}_2| = k_2$, respectively. Clearly, the supermodel with lower dimension, $\mathcal{M}_{\mathbf{s}_1}$, is closer to the true model, and the theory suggests that the Bayes factor $\log BF_{\mathbf{s}_2,\mathbf{s}_1}$ decays with growing n. The linear

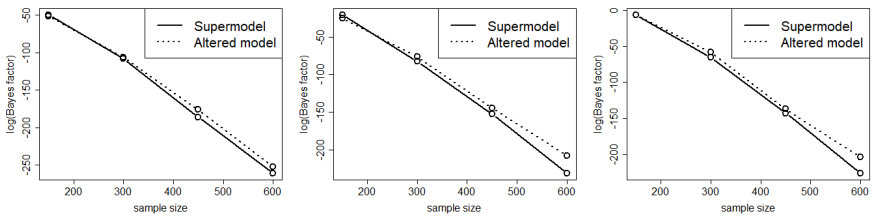


Figure S-3: Line diagram showing how the Bayes factors decrease as sample sizes increase in autoregressive model when $|\mathbf{s}_0| = 10$ (left panel), $|\mathbf{s}_0| = 40$ (middle panel), and $|\mathbf{s}_0| = 40$ (right panel)

regression and Gaussian process regression with squared exponential kernel is considered.

In the linear regression, we choose $k_1 = 1$ and $k_2 = 5$. Everything else is kept same as in Section S-1.1, expect here we choose $\boldsymbol{\beta}_{0,\mathbf{s}} = (1/2, \ldots, 1/2)$. In the Gaussian process regression, we choose $k_1 = 5$ and $k_2 = 15$. Everything is kept same as in Section S-1.2. The results are summarized in Figure S-4.

Observe that for both the cases we observe a sharp linear decrease of log Bayes factors as n increases, which validates our theoretical results.

S-2 A generic TTMCMC sampler for variable selection

Here we devise a novel TTMCMC algorithm for generic variable selection problems using mixtures of additive and multiplicative transformations of singleton variables, further supplementing with a deterministic transformation step to enhance mixing. Given a set of existing covariates, we propose a new covariate in the "birth move" by Bayes Information Criterion (BIC). We compute Bayes factors from the available TTMCMC realizations to compare subsets of the covariates. Interestingly, the acceptance ratios of neither TMCMC, nor TTMCMC, depend upon the proposal distributions, even if they are not symmetric, and even for dimension-changing moves. Thus, these approaches are novel compared to the traditional fixed-dimensional Metropolis-Hastings and the variable-dimensional RJMCMC approach.

We provide our general TTMCMC sampler for variable selection in the form of Algorithm S-2.1. We assume that $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\vartheta})$ is the set of parameters

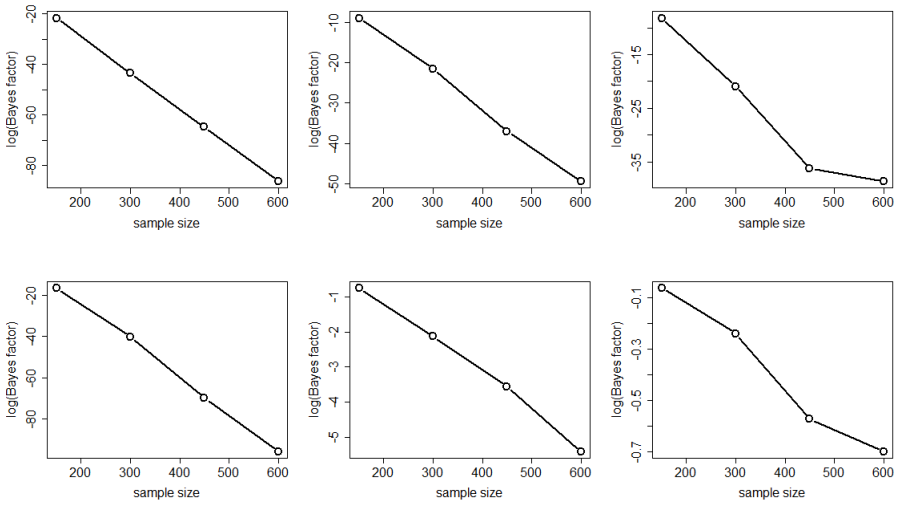


Figure S-4: Line diagram showing how the Bayes factor favors the better model as sample sizes increase in the misspecified models setup, in linear regression (top row), and Gaussian process regression (bottom row), when $|\mathbf{s}_0| = 10$ (left panel), $|\mathbf{s}_0| = 40$ (middle panel), and $|\mathbf{s}_0| = 70$ (right panel).

associated with the model, $\boldsymbol{\beta}$ being the k-dimensional regression coefficients associated with the chosen covariates, where k is a random variable. The parameter vector $\boldsymbol{\vartheta}$ consists of other sets of parameters, and may even contain several other parameter vectors associated with the covariates, having the same (variable) dimension k as $\boldsymbol{\beta}$. For instance, in a Gaussian process regression, the mean function may be modeled by a linear regression with regression coefficients $\boldsymbol{\beta}$ and the covariance function may be modeled by a squared exponential kernel consisting of smoothness parameters having the same random dimension k as $\boldsymbol{\beta}$. We shall denote by $\pi(\boldsymbol{\theta}, \mathbf{s}, k)$ as proportional to the product of the prior and the likelihood, where $\mathbf{s}$ and k, the random subset of covariate indices and its cardinality, are also considered unknown and suitable priors are envisaged for the same. Thus, with abuse of notation for convenience and simplicity, we write the posterior $\pi(\boldsymbol{\theta}, k)$ as

$$\pi(\boldsymbol{\theta}, \mathbf{s}, k) \propto L(\boldsymbol{\theta}|\mathbf{s}, k)\pi(\boldsymbol{\theta}|\mathbf{s}, k)\pi(\mathbf{s}|k)\pi(k), \tag{S-1}$$

where $\pi(k)$ denotes the prior for k, $\pi(\boldsymbol{\theta}|\mathbf{s}, k)$ stands for the prior for $\boldsymbol{\theta}$ given $\mathbf{s}$ and k, $\pi(\mathbf{s}|k)$ is the prior for $\mathbf{s}$ given k and $L(\boldsymbol{\theta}|\mathbf{s}, k)$ is the likelihood for $\boldsymbol{\theta}$ given $\mathbf{s}$ and k. Given k, we set the uniform prior for $\mathbf{s}$:

$$\pi(\mathbf{s}|k) = \frac{1}{\binom{p}{k}}, \text{ for } k = 1, \ldots, p. \tag{S-2}$$

Algorithm S-2.1. *General TTMCMC algorithm for variable selection.*

- Let the initial value be $\boldsymbol{\theta}^{(0)} = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\vartheta}^{(0)})$, where $\boldsymbol{\beta}^{(0)} \in \mathbb{R}^{k^{(0)}}$, are the coefficients of the $k^{(0)}$ covariates in the current regression model, and $\boldsymbol{\vartheta}^{(0)}$ consists of the initial values of the other model parameters, which may even include other $k^{(0)}$-dimensional parameters associated with the covariates in the model. Also let $\mathbf{s}^{(0)}$ denote the initial choice for the subset of indices for the covariates associated with the model.

- For $t = 0, 1, 2, \ldots$

  1. Generate $u = (u_1, u_2, u_3) \sim Multinomial(1; w_{b,k^{(t)}}, w_{d,k^{(t)}}, w_{nc,k^{(t)}})$, where $w_{b,k^{(t)}}, w_{d,k^{(t)}}, w_{nc,k^{(t)}}$ are birth, death and no-change probabilities, given $k^{(t)}$. Hence, $w_{b,k^{(t)}}, w_{d,k^{(t)}}, w_{nc,k^{(t)}}$ are non-negative and $w_{b,k^{(t)}} + w_{d,k^{(t)}} + w_{nc,k^{(t)}} = 1$. Also, $w_{b,k^{(t)}} = 0$ if $k^{(t)} = |\mathbf{S}|$ and $w_{d,k^{(t)}} = 0$ if $k^{(t)} = 1$.

2. If $u_1 = 1$ (increase dimension by selecting a new covariate), generate $U \sim U(0,1)$ and do the following:

   (a) If $U \leq \tilde{p}$, where $\tilde{p} \in [0,1]$ (use additive transformation for dimension change),

      i. Given $\mathbf{s}^{(t)}$, the current subset of covariates and the current set of parameters $\boldsymbol{\theta}^{(t)}$, select a new covariate $\{x_{ir} : i = 1, \ldots, n\}$, where $r \in \mathbf{S} \backslash \mathbf{s}^{(t)}$, by minimizing $BIC(u)$, for $u \in \mathbf{S} \backslash \mathbf{s}^{(t)}$. Here $BIC(u)$ stands for the BIC when the model consists of the covariates indexed by $\{\mathbf{s}^{(t)}, u\}$. Let $\mathbf{s}' = \{\mathbf{s}^{(t)}, r\}$.

      ii. Randomly select a co-ordinate from $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \ldots, \beta_{k^{(t)}}^{(t)})$ assuming uniform probability $1/k^{(t)}$ for each co-ordinate. Let j denote the chosen co-ordinate.

      iii. Generate $\epsilon_1 \sim N(0,1)$ and propose the following birth move:

$$\boldsymbol{\beta}' = (\beta_1^{(t)}, \ldots, \beta_{j-1}^{(t)}, \beta_j^{(t)} + a_{\beta,j}|\epsilon_1|, \beta_j^{(t)} - a_{\beta,j}|\epsilon_1|, \beta_{j+1}^{(t)}, \ldots, \beta_{k^{(t)}}^{(t)}).$$

      Here $a_{\beta,j}$ is the appropriate positive scaling constant associated with the j-th co-ordinate of $\boldsymbol{\beta}$. In general, $a_{\theta,j}$ will stand for the appropriate positive scaling constant associated with the j-th co-ordinate of $\boldsymbol{\theta}$.

      iv. Re-label the elements of $\boldsymbol{\beta}'$ as $(\beta_1', \beta_2', \ldots, \beta_{k^{(t)}+1}')$.

         A. If there is another set of real-valued variable-dimensional parameters, say, $\boldsymbol{\gamma}$, associated with the covariates, then also generate $\epsilon_2 \sim N(0,1)$ and propose

$$\boldsymbol{\gamma}' = (\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j^{(t)} + a_{\gamma,j}|\epsilon_2|, \gamma_j^{(t)} - a_{\gamma,j}|\epsilon_2|, \gamma_{j+1}^{(t)}, \ldots, \gamma_{k^{(t)}}^{(t)}).$$

         B. Re-label the elements of $\boldsymbol{\gamma}'$ as $(\gamma_1', \gamma_2', \ldots, \gamma_{k^{(t)}+1}')$.

         C. Repeat the procedure for further sets of variable-dimensional parameters related to the covariates.

         D. Keep all other elements of $\boldsymbol{\theta}$ unchanged, and refer to the entire set of proposed parameter values as $\boldsymbol{\theta}'$.

      v. If $\boldsymbol{\beta}$ is the only variable-dimensional parameter related to the covariates, then the acceptance probability

of the birth move is:

$$a_b = \min\left\{1, \frac{1}{k^{(t)}+1} \times \frac{w_{d,k^{(t)}+1}}{w_{b,k^{(t)}}} \ \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times 2a_{\beta,j}\right\}.$$

A. If $\boldsymbol{\gamma}$ is another real-valued variable-dimensional parameter related to the covariates, then the acceptance probability of the birth move is:

$$a_b = \min\left\{1, \frac{1}{k^{(t)}+1} \times \frac{w_{d,k^{(t)}+1}}{w_{b,k^{(t)}}} \ \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times 2a_{\beta,j} \times 2a_{\gamma,j}\right\},$$

that is, $2a_{\gamma,j}$ must also be multiplied to the acceptance ratio.

B. For further real-valued variable-dimensional parameter associated with the covariates, the process must be continued by further multiplying twice the scaling constant of the relevant parameter to the acceptance ratio.

vi. Set

$$(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = \begin{cases} (\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1) & \text{with probability} \quad a_b \\ (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}) & \text{with probability} \quad 1-a_b. \end{cases}$$

(b) If $U > \tilde{p}$ (use multiplicative transformation for dimension change),

i. Given $\mathbf{s}^{(t)}$, the current subset of covariates and the current set of parameters $\boldsymbol{\theta}^{(t)}$, select a new covariate $\{x_{ir} : i = 1, \ldots, n\}$, where $r \in \mathbf{S} \backslash \mathbf{s}^{(t)}$, by minimizing $BIC(u)$, for $u \in \mathbf{S} \backslash \mathbf{s}^{(t)}$. Let $\mathbf{s}' = \{\mathbf{s}^{(t)}, r\}$.

ii. Randomly select a co-ordinate from $\boldsymbol{\beta}^{(t)} = (\beta_1^{(t)}, \ldots, \beta_{k^{(t)}}^{(t)})$ assuming uniform probability $1/k^{(t)}$ for each co-ordinate. Let j denote the chosen co-ordinate.

iii. Generate $\epsilon_1 \sim U(-1, 1)$ and propose the following birth move:

$$\boldsymbol{\beta}' = (\beta_1^{(t)}, \ldots, \beta_{j-1}^{(t)}, \beta_j^{(t)}\epsilon_1, \beta_j^{(t)}/\epsilon_1, \beta_{j+1}^{(t)}, \ldots, \beta_{k^{(t)}}^{(t)}).$$

iv. Re-label the elements of $\boldsymbol{\beta}'$ as $(\beta_1', \beta_2', \ldots, \beta_{k^{(t)}+1}')$.

A. If there is another set of real-valued variable-dimensional parameters, say, $\boldsymbol{\gamma}$, associated with the covariates, then also generate $\epsilon_2 \sim U(-1, 1)$ and propose

$$\boldsymbol{\gamma}' = (\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j^{(t)}\epsilon_2, \gamma_j^{(t)}/\epsilon_2, \gamma_{j+1}^{(t)}, \ldots, \gamma_{k^{(t)}}^{(t)}).$$

B. Re-label the elements of $\boldsymbol{\gamma}'$ as $(\gamma'_1, \gamma'_2, \ldots, \gamma'_{k^{(t)}+1})$.

C. Repeat the procedure for further sets of variable-dimensional parameters related to the covariates.

D. Keep all other elements of $\boldsymbol{\theta}$ unchanged, and refer to the entire set of proposed parameter values as $\boldsymbol{\theta}'$.

v. If $\boldsymbol{\beta}$ is the only variable-dimensional parameter related to the covariates, then the acceptance probability of the birth move is:

$$a_b = \min\left\{1, \frac{1}{k^{(t)}+1} \times \frac{w_{d,k^{(t)}+1}}{w_{b,k^{(t)}}} \times \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times \frac{|\beta_j^{(t)}|}{|\epsilon_1|}\right\}.$$

A. If $\boldsymbol{\gamma}$ is another real-valued variable-dimensional parameter related to the covariates, then the acceptance probability of the birth move is:

$$a_b = \min\left\{1, \frac{1}{k^{(t)}+1} \times \frac{w_{d,k^{(t)}+1}}{w_{b,k^{(t)}}} \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times \frac{|\beta_j^{(t)}|}{|\epsilon_1|} \times \frac{|\gamma_j^{(t)}|}{|\epsilon_2|}\right\}.$$

B. For further variable-dimensional parameter associated with the covariates, noting that the process must be continued by further multiplying the ratio of the absolute value of the current parameter value and the relevant ϵ, to the acceptance ratio.

vi. Set

$$(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = \begin{cases} (\boldsymbol{\theta}', \mathbf{s}', k^{(t)}+1) & \text{with probability} \quad a_b \\ (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}) & \text{with probability} \quad 1-a_b. \end{cases}$$

3. If $u_2 = 1$ (decrease dimension by deleting an existing covariate), generate $U \sim U(0, 1)$ and do the following:

(a) If $U \leq \tilde{p}$ (use additive transformation for dimension change),

i. Randomly select a co-ordinate j from $\{1,\ldots,k^{(t)}\}$ assuming uniform probability $1/k^{(t)}$ for each co-ordinate, and randomly select a co-ordinate j' from $\{1,\ldots,k^{(t)}\}\backslash\{j\}$ with probability $1/(k^{(t)}-1)$. Assuming $j<j'$, let $\beta_j^*=(\beta_j^{(t)}+\beta_{j'}^{(t)})/2$. Replace $\beta_j^{(t)}$ with β_j^* and delete $\beta_{j'}^{(t)}$.

ii. Delete $\{x_{ij'}:i=1,\ldots,n\}$. Let $\mathbf{s}'=\mathbf{s}^{(t)}\backslash\{j'\}$.

iii. Propose the following death move:

$$\boldsymbol{\beta}'=(\beta_1^{(t)},\ldots,\beta_{j-1}^{(t)},\beta_j^*,\beta_{j+1}^{(t)},\ldots,\beta_{j'-1}^{(t)},\beta_{j'+1}^{(t)},\ldots,\beta_{k^{(t)}}^{(t)}).$$

iv. Re-label the elements of $\boldsymbol{\beta}'$ as $(\beta_1',\beta_2',\ldots,\beta_{k^{(t)}-1}')$.

A. If there is another set of real-valued variable-dimensional parameters, say, $\boldsymbol{\gamma}$, associated with the covariates, then propose

$$\boldsymbol{\gamma}'=(\gamma_1^{(t)},\ldots,\gamma_{j-1}^{(t)},\gamma_j^*,\gamma_{j+1}^{(t)},\ldots,\gamma_{j'-1}^{(t)},\gamma_{j'+1}^{(t)},\ldots,\gamma_{k^{(t)}}^{(t)}),$$

where $\gamma_j^*=(\gamma_j^{(t)}+\gamma_{j'}^{(t)})/2$.

B. Re-label the elements of $\boldsymbol{\gamma}'$ as $(\gamma_1',\gamma_2',\ldots,\gamma_{k^{(t)}-1}')$.

C. Repeat the procedure for further sets of variable-dimensional parameters related to the covariates.

D. Keep all other elements of $\boldsymbol{\theta}$ unchanged, and refer to the entire set of proposed parameter values as $\boldsymbol{\theta}'$.

v. If $\boldsymbol{\beta}$ is the only variable-dimensional parameter related to the covariates, then the acceptance probability of the death move is:

$$a_d=\min\left\{1,k^{(t)}\times\frac{w_{b,k^{(t)}-1}}{w_{d,k^{(t)}}}\ \frac{\pi\left(\boldsymbol{\theta}',\mathbf{s}',k^{(t)}-1\right)}{\pi\left(\boldsymbol{\theta}^{(t)},\mathbf{s}^{(t)},k^{(t)}\right)}\times\frac{1}{2a_{\beta,j}}\right\}.$$

A. If $\boldsymbol{\gamma}$ is another real-valued variable-dimensional parameter related to the covariates, then the acceptance probability of the death move is:

$$a_d=\min\left\{1,k^{(t)}\times\frac{w_{b,k^{(t)}-1}}{w_{d,k^{(t)}}}\ \frac{\pi\left(\boldsymbol{\theta}',\mathbf{s}',k^{(t)}-1\right)}{\pi\left(\boldsymbol{\theta}^{(t)},\mathbf{s}^{(t)},k^{(t)}\right)}\times\frac{1}{2a_{\beta,j}}\times\frac{1}{2a_{\gamma,j}}\right\},$$

that is, $1/(2a_{\gamma,j})$ must also be multiplied to the acceptance ratio.

B. For further real-valued variable-dimensional parameter associated with the covariates, the process must be continued in the above manner.

vi. Set

$$(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = \begin{cases} (\boldsymbol{\theta}', \mathbf{s}', k^{(t)} - 1) & \text{with probability} \quad a_d \\ (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}) & \text{with probability} \quad 1 - a_d. \end{cases}$$

(b) If $U > \tilde{p}$ (use multiplicative transformation for dimension change),

i. Randomly select a co-ordinate j from $\{1, \ldots, k^{(t)}\}$ assuming uniform probability $1/k^{(t)}$ for each co-ordinate, and randomly select a co-ordinate j' from $\{1, \ldots, k^{(t)}\} \backslash \{j\}$ with probability $1/(k^{(t)}-1)$. Assuming $j < j'$, let $\beta_j^* = \sqrt{|\beta_j^{(t)} \beta_{j'}^{(t)}|}$ with probability $1/2$ and set $\beta_j^* = -\sqrt{|\beta_j^{(t)} \beta_{j'}^{(t)}|}$ with the remaining probability. Replace $\beta_j^{(t)}$ with β_j^* and delete $\beta_{j'}^{(t)}$.

ii. Delete $\{x_{ij'} : i = 1, \ldots, n\}$. Let $\mathbf{s}' = \mathbf{s}^{(t)} \backslash \{j'\}$.

iii. Propose the following death move:

$$\boldsymbol{\beta}' = (\beta_1^{(t)}, \ldots, \beta_{j-1}^{(t)}, \beta_j^*, \beta_{j+1}^{(t)}, \ldots, \beta_{j'-1}^{(t)}, \beta_{j'+1}^{(t)}, \ldots, \beta_{k^{(t)}}^{(t)}).$$

iv. Re-label the elements of $\boldsymbol{\beta}'$ as $(\beta_1', \beta_2', \ldots, \beta_{k^{(t)}-1}')$.

A. If there is another set of real-valued variable-dimensional parameters, say, $\boldsymbol{\gamma}$, associated with the covariates, then propose

$$\boldsymbol{\gamma}' = (\gamma_1^{(t)}, \ldots, \gamma_{j-1}^{(t)}, \gamma_j^*, \gamma_{j+1}^{(t)}, \ldots, \gamma_{j'-1}^{(t)}, \gamma_{j'+1}^{(t)}, \ldots, \gamma_{k^{(t)}}^{(t)}),$$

where $\gamma_j^* = \sqrt{|\gamma_j^{(t)} \gamma_{j'}^{(t)}|}$ or $-\sqrt{|\gamma_j^{(t)} \gamma_{j'}^{(t)}|}$ with equal probabilities.

B. Re-label the elements of $\boldsymbol{\gamma}'$ as $(\gamma_1', \gamma_2', \ldots, \gamma_{k^{(t)}-1}')$.

C. Repeat the procedure for further sets of variable-dimensional parameters related to the covariates.

D. Keep all other elements of $\boldsymbol{\theta}$ unchanged, and refer to the entire set of proposed parameter values as $\boldsymbol{\theta}'$.

v. If $\boldsymbol{\beta}$ is the only variable-dimensional parameter related to the covariates, then the acceptance probability of the death move is:

$$a_d = \min\left\{1, k^{(t)} \times \frac{w_{b,k^{(t)}-1}}{w_{d,k^{(t)}}} \; \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}-1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times \frac{1}{|\beta_{j'}^{(t)}|}\right\}.$$

A. If $\boldsymbol{\gamma}$ is another real-valued variable-dimensional parameter related to the covariates, then the acceptance probability of the death move is:

$$a_d = \min\left\{1, k^{(t)} \times \frac{w_{b,k^{(t)}-1}}{w_{d,k^{(t)}}} \; \frac{\pi\left(\boldsymbol{\theta}', \mathbf{s}', k^{(t)}-1\right)}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)} \times \frac{1}{|\beta_{j'}^{(t)}|} \times \frac{1}{|\gamma_{j'}^{(t)}|}\right\},$$

that is, $1/|\gamma_{j'}^{(t)}|$ must also be multiplied to the acceptance ratio.

B. For further real-valued variable-dimensional parameter associated with the covariates, the process must be continued in the above manner.

vi. Set

$$(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = \begin{cases} (\boldsymbol{\theta}', \mathbf{s}', k^{(t)}-1) & \text{with probability} \quad a_d \\ (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}) & \text{with probability} \quad 1-a_d. \end{cases}$$

4. If $u_3 = 1$ (dimension remains unchanged), then given that there are d dimensions in the current iteration, generate $U \sim U(0,1)$.

(a) If $U \leq \tilde{p}$, then do the following:

(i) For parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, etc. associated with the covariates, for $j = 1, \ldots, k^{(t)}$, set $\tilde{a}_{\beta,j} = c a_{\beta,j}$, $\tilde{a}_{\gamma,j} = c a_{\gamma,j}$, etc. where $c \in (0,1)$ is some appropriate constant. For all other parameter co-ordinates θ_j, let $\tilde{a}_{\theta,j} = a_{\theta,j}$.

(ii) Generate $\varepsilon \sim N(0,1)$, $b_j \overset{iid}{\sim} U(\{-1,1\})$ for $j = 1, \ldots, d$, and set $\theta_j' = \theta_j^{(t)} + b_j \tilde{a}_{\theta,j} |\varepsilon|$, for $j = 1, \ldots, d$.

(iii) Evaluate

$$\alpha_1 = \min\left\{1, \frac{\pi(\boldsymbol{\theta}', \mathbf{s}^{(t)}, k^{(t)})}{\pi\left(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)}\right)}\right\}.$$

(iv) Set $(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}', \mathbf{s}^{(t)}, k^{(t)})$ with probability α_1, else set $(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)})$.

(b) If $U > \tilde{p}$, then do the following:

(i) Generate $\varepsilon \sim U(-1, 1)$, $b_j \overset{iid}{\sim} U(\{-1, 0, 1\})$ for $j = 1, \ldots, d$, and set $\theta'_j = \theta_j^{(t)}\varepsilon$ if $b_j = 1$, $\theta'_j = \theta_j^{(t)}/\varepsilon$ if $b_j = -1$ and $\theta'_j = \theta_j^{(t)}$ if $b_j = 0$, for $j = 1, \ldots, d$. Calculate $|J| = |\varepsilon|^{\sum_{j=1}^{d} b_j}$.

(ii) Evaluate

$$\alpha_2 = \min\left\{1, \frac{\pi(\boldsymbol{\theta}', \mathbf{s}^{(t)}, k^{(t)})}{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)})} \times |J|\right\}.$$

(iii) Set $(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}', \mathbf{s}^{(t)}, k^{(t)})$ with probability α_2, else set $(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k^{(t)})$.

5. (*Mixing-enhancement step*) Assume that there are d dimensions in the current iteration after implementing either of the birth, death and no-change steps. Generate $U \sim U(0, 1)$.

(a) If $U \leq \tilde{q}$, where $\tilde{q} \in (0, 1)$, then do the following

(i) For parameters $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, etc. associated with the covariates, for $j = 1, \ldots, k^{(t+1)}$, set $\tilde{a}_{\beta,j} = ca_{\beta,j}$, $\tilde{a}_{\gamma,j} = ca_{\gamma,j}$, etc. where $c \in (0, 1)$ is some appropriate constant. For all other parameter co-ordinates θ_j, let $\tilde{a}_{\theta,j} = a_{\theta,j}$.

(ii) Generate $\tilde{U} \sim U(0, 1)$ and $\varepsilon \sim N(0, 1)$. If $\tilde{U} < 1/2$, set $\theta''_j = \theta_j^{(t+1)} + \tilde{a}_{\theta,j}|\varepsilon|$, for $j = 1, \ldots, d$; else, set $\theta''_j = \theta_j^{(t+1)} - \tilde{a}_{\theta,j}|\varepsilon|$, for $j = 1, \ldots, d$.

(iii) Letting $\boldsymbol{\theta}'' = (\theta''_1, \ldots, \theta''_d)$, evaluate

$$\alpha_3 = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'', \mathbf{s}^{(t+1)}, k^{(t+1)})}{\pi(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)})}\right\}.$$

(iv) Set $(\tilde{\boldsymbol{\theta}}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}'', \mathbf{s}^{(t+1)}, k^{(t+1)})$ with probability α_3, else set $(\tilde{\boldsymbol{\theta}}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)})$.

(b) If $U > \tilde{q}$, then

(i) Generate $\varepsilon \sim U(-1, 1)$ and $\tilde{U} \sim U(0, 1)$. If $\tilde{U} < 1/2$, set $\theta''_j = \theta_j^{(t+1)}\varepsilon$ for $j = 1, \ldots, d$ and $|J| = |\varepsilon|^d$,

else set $\theta''_j = \theta_j^{(t+1)}/\varepsilon$ for $j = 1, \ldots, d$ and $|J| = |\varepsilon|^{-d}$.

(ii) Evaluate

$$\alpha_4 = \min\left\{1, \frac{\pi(\boldsymbol{\theta}'', \mathbf{s}^{(t+1)}, k^{(t+1)})}{\pi(\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)})} \times |J|\right\}.$$

(iii) Set $(\tilde{\boldsymbol{\theta}}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}'', \mathbf{s}^{(t+1)}, k^{(t+1)})$ with probability α_4, else set $(\tilde{\boldsymbol{\theta}}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)}) = (\boldsymbol{\theta}^{(t+1)}, \mathbf{s}^{(t+1)}, k^{(t+1)})$.

- End for
- Store $\{(\tilde{\boldsymbol{\theta}}^{(0)}, \mathbf{s}^{(0)}, k^{(0)}), (\tilde{\boldsymbol{\theta}}^{(1)}, \mathbf{s}^{(1)}, k^{(1)}), \ldots\}$ for Bayesian inference.

The main strategies proposed in the general TTMCMC Algorithm S-2.1 for variable selection require some elucidation. In this regard, a few remarks are in order.

First, we propose a mixture of additive and multiplicative transformations in all the steps of the algorithm, since it has been observed in Dey and Bhattacharya (2016) that such mixture proposal induces better mixing that either additive or multiplicative transformations using the localised moves of the additive transformation and the non-localised ("random dive") moves of the multiplicative transformation (see also Dutta (2012) for some theoretical details on random dive).

In the dimension-changing steps 2. and 3. of Algorithm S-2.1, except for the parameters associated with increase or decrease of the dimension, we have proposed to keep all the remaining parameters fixed. Fixing the other parameters is not necessary for the validity of TTMCMC; indeed, Das and Bhattacharya (2019) proposed to update all the parameters even in the dimension-changing steps. However, in our variable selection experiments, fixing the remaining parameters led to significantly improved acceptance rates of the birth and death steps compared to the strategy of updating all the unknowns simultaneously. The choice of the positive scales $a_{\theta,j}$ in the additive transformation part plays important role here. To elucidate, note that it is natural to expect high acceptance rates with sufficiently small scales in fixed-dimensional problems, but in our variable-dimensional setup, observe that the acceptance ratios for the birth and death steps depend upon the scales of the parameters selected for birth and death. If the scales are generally chosen to be small, then the acceptance rate for the birth move

would be small as well. On the other hand, if the scales are generally chosen to be relatively large, then the acceptance rate for the entire dimension-changing move would be small, for a relatively large number of parameters. With these small or large scale choices, the acceptance ratios in the no-change (fixed-dimensional) step 4. and the mixing-enhancement step 5. would also be small.

We attempt to solve all the above problems with the strategy of choosing somewhat large scales $a_{\theta,j}$ and by fixing the parameters in the birth and death steps that are not involved in dimension-change. The relatively large scales would ensure adequate acceptance rate for the birth move; note that the scales should not be so large as to reduce the death rate significantly. Now, these large scales would also diminish the acceptance rates in the no-change and the mixing-enhancement steps. To counter this, we multiply the scales of the parameters associated with the covariates by $c \in (0, 1)$ in those steps, which is a valid mathematical strategy in the sense of satisfying detailed balance. Further discussion regarding these will be provided in course of the applications of Algorithm S-2.1.

The fixed-dimensional mixing-enhancement step has parallels with Liu and Sabatti (2000) (see also the supplement of Dutta and Bhattacharya (2014) and Algorithm 2 of Roy and Bhattacharya (2020)). Indeed, it has been observed that the strategy can often drastically improve the mixing properties in fixed-dimensional setups.

Finally, note that $\mathbf{s}$ and k are not updated in the no-change and mixing enhancing steps, so that $\pi(\mathbf{s}|k)\pi(k)$ gets cancelled in the corresponding acceptance ratios.

S-3 Bayes factor computation using TTMCMC realizations

Assuming that there are N realizations of TTMCMC stored for Bayesian inference after discarding a suitable burn-in period, the Bayes factors associated with the distinct subsets of the covariates featuring in the TTMCMC samples can be calculated as follows.

Let there be $\tilde{N}$ $(< N)$ distinct subsets $\left\{\mathbf{s}_1^*, \mathbf{s}_2^*, \ldots, \mathbf{s}_{\tilde{N}}^*\right\}$ in the TTMCMC sample, each subset consisting of distinct indices of a set of covariates which is a subset of the entire pool of covariates indexed by $\mathbf{S}$. Thus, the TTMCMC sample consists of $\tilde{N}$ distinct subsets of covariates out of a total $2^p - 1$ possibilities, $p = |\mathbf{S}|$ being the total available number of covariates. The subsets of covariates that did not feature in the TTMCMC sample will be

interpreted as having negligible posterior probabilities and will be not be considered any further for our Bayesian analyses.

For $i = 1, \ldots, \tilde{N}$, assuming that $\mathbf{s}_i^*$ is repeated N_i times in the TTMCMC sample, so that $\sum_{i=1}^{\tilde{N}} N_i = N$, we estimate its posterior probability by $\tilde{\pi}(\mathbf{s}_i^*) = N_i/N$. Let $k_i^* = |\mathbf{s}_i^*|$ be the cardinality of $\mathbf{s}_i^*$. Note that the prior for the model associated with any subset $\mathbf{s}$ consisting of k covariates is uniform over all $\binom{p}{k}$ possibilities, given by (S-2). Hence, the marginal prior probability of $\mathbf{s}$ with $|\mathbf{s}| = k$ is

$$\pi(\mathbf{s}) = \sum_{j=1}^{p} \pi(\mathbf{s}|j)\pi(j) = \pi(\mathbf{s}|k)\pi(k) = \frac{\pi(k)}{\binom{p}{k}}, \tag{S-1}$$

since $\pi(\mathbf{s}|j) = 0$ if $j \neq k$. In the above, $\pi(k)$ denotes the prior for k.

Using (S-1), we compute for each $i = 1, \ldots, \tilde{N}$,

$$B_i = \frac{\tilde{\pi}(\mathbf{s}_i^*)}{\pi(\mathbf{s}_i^*)} = \frac{N_i}{N} \times \frac{\binom{p}{k_i^*}}{\pi(k_i^*)}. \tag{S-2}$$

For any $i, j \in \{1, \ldots, \tilde{N}\}$, the (approximate) Bayes factor of the model associated with $\mathbf{s}_i^*$ against that associated with $\mathbf{s}_j^*$ is given by

$$BF_{ij} = B_i/B_j. \tag{S-3}$$

Thus, the best model is the one with the largest B_i; $i = 1, \ldots, \tilde{N}$. Note that B_i is proportional to the marginal density of the data, given the i-th model, where the proportionality constant is the same for all the competing models.

S-4 Proof of convergence of the TTMCMC algorithm

To prove convergence of Algorithm S-2.1 it is sufficient to establish detailed balance, irreducibility and aperiodicity of the algorithm, which we undertake step-by-step in this section. For simplicity, let us assume that $\boldsymbol{\beta}$ is the only parameter vector associated with the covariates. The extension is trivial for other parameter vectors associated with the covariates.

S-4.1 Proof of detailed balance

S-4.1.1 Additive transformation

Let us first consider the case of the additive transformation, which we select with probability $\tilde{p}$. To see that detailed balance is satisfied for the birth

and death moves, note that associated with the birth move, the probability (essentially) of transition $(\boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)}, k) \mapsto (\boldsymbol{\beta}', \mathbf{s}', k+1)$, with $k = |\mathbf{s}^{(t)}|$ and $k+1 = |\mathbf{s}'|$ (so that $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^k$ and $\boldsymbol{\beta}' \in \mathbb{R}^{k+1}$), while the other elements of $\boldsymbol{\theta}$ are held fixed, is given by:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \times \tilde{p} \times \frac{1}{k} \times w_{b,k} \times N(\epsilon : 0, 1) \\
&\qquad \times \min\left\{1, \frac{1}{k+1} \times \frac{w_{d,k+1}}{w_{b,k}} \times \frac{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1)}{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k)} \times \left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}, \epsilon)}\right|\right\} \\
&= \tilde{p} \times N(\epsilon : 0, 1) \times \min\left\{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k+1) \times \frac{1}{k} \times w_{b,k},\right. \\
&\qquad \left.\frac{1}{k(k+1)} \times w_{d,k+1} \times \pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times \left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon)}\right|\right\}, \qquad \text{(S-1)}
\end{aligned}$$

where $N(\epsilon : 0, 1)$ is the density of the normal distribution with mean 0 and variance 1, evaluated at ϵ. Assuming that $\beta_j^{(t)}$ was selected, and was split into $\beta_j^{(t)} + a_{\beta,j}\epsilon$ and $\beta_j^{(t)} - a_{\beta,j}\epsilon$, $\left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}, \epsilon)}\right| = 2a_{\beta,j}$.

At the reverse death move we must be able to return to $(\boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)}, k)$ from $(\boldsymbol{\beta}', \mathbf{s}', k+1)$, while the other elements of $\boldsymbol{\theta}$ are held fixed. We select β_j' with probability $1/(k+1)$, then select β_{j+1}' without replacement with probability $1/k$, and take the resultant average.

Let ϵ^* be such that $\beta_j^{(t)} + a_{\beta,j}\epsilon^* = \beta_j'$ and $\beta_j^{(t)} - a_{\beta,j}\epsilon^* = \beta_{j+1}'$, so that $\epsilon^* = (\beta_j' - \beta_{j+1}')/2$. The transition probability of the death move is hence given by:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times \tilde{p} \times w_{d,k+1} \times N(\epsilon, 0, 1) \times \frac{1}{k+1} \times \frac{1}{k} \times \left|\frac{\partial(\boldsymbol{\beta}', \epsilon)}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)}\right| \\
&\qquad \times \min\left\{1, (k+1) \times \frac{w_{b,k}}{w_{d,k+1}} \times \frac{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k)}{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1)} \times \left|\frac{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)}{\partial(\boldsymbol{\beta}', \epsilon)}\right|\right\} \\
&= \tilde{p} \times N(\epsilon : 0, 1) \times \min\left\{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times w_{d,k+1} \times \frac{1}{k(k+1)} \times \left|\frac{\partial(\boldsymbol{\beta}', \epsilon)}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)}\right|,\right. \\
&\qquad \left.\frac{1}{k} \times w_{b,k} \times \pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k)\right\} \\
&= \tilde{p} \times N(\epsilon : 0, 1) \times \min\left\{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times w_{d,k+1} \times \frac{1}{k(k+1)} \times 2a_{\beta,j},\right.
\end{aligned}$$

$$\left. \frac{1}{k} \times w_{b,k} \times \pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \right\}. \tag{S-2}$$

Thus, (S-1) = (S-2), showing that detailed balance holds for the birth and the death moves. The proof of detailed balance for the no-change move type where the dimension remains unchanged is the same as that of TMCMC, and has been been proved in the supplement of Dutta and Bhattacharya (2014).

S-4.1.2 Multiplicative transformation

Now let us consider the multiplicative transformation, which we select with probability $1-\tilde{p}$. For the birth move, the probability (essentially) of the transition $(\boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)}, k) \mapsto (\boldsymbol{\beta}', \mathbf{s}, k+1))$, while the other elements of $\boldsymbol{\theta}$ are held fixed, is given by:

$$\begin{aligned}
&\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \times (1-\tilde{p}) \times \frac{1}{k} \times w_{b,k} \times U(\epsilon : -1, 1) \\
&\qquad \times \min\left\{1, \frac{1}{k+1} \times \frac{w_{d,k+1}}{w_{b,k}} \times \frac{1}{2} \times \frac{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1)}{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k)} \times \left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon)}\right|\right\} \\
&= (1-\tilde{p}) \times U(\epsilon : -1, 1) \times \min\left\{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \times \frac{1}{k} \times w_{b,k}, \right. \\
&\qquad \left. \frac{1}{k(k+1)} \times w_{d,k+1} \times \pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times \frac{1}{2} \times \left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon)}\right|\right\},
\end{aligned} \tag{S-3}$$

where $U(\epsilon : -1, 1)$ is the density of the uniform distribution on $[-1, 1]$, evaluated at ϵ. Assuming that $\beta_j^{(t)}$ was selected, and was split into $\beta_j^{(t)}\epsilon$ and $\beta_j^{(t)}/\epsilon$, $\left|\frac{\partial \boldsymbol{\beta}'}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon)}\right| = 2|\beta_j^{(t)}|/|\epsilon|$.

At the reverse death move we must be able to return to $(\boldsymbol{\beta}^{(t)}, \mathbf{s}^{(t)}, k)$ from $(\boldsymbol{\beta}', \mathbf{s}', k+1)$, while the other elements of $\boldsymbol{\theta}$ are held fixed. We select β_j' with probability $1/(k+1)$, then select β_{j+1}' without replacement with probability $1/k$, and take $\sqrt{|\beta_j'\beta_{j+1}'|})$ or $-\sqrt{|\beta_j'\beta_{j+1}'|})$ with equal probabilities.

Let ϵ^* be such that $\beta_j^{(t)}\epsilon^* = \beta_j'$ and $\beta_j^{(t)}/\epsilon^* = \beta_{j+1}'$, so that $\epsilon^* = \pm\sqrt{|\beta_j'\beta_{j+1}'|})$. The transition probability of the death move is hence given

by:

$$\pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times (1-\tilde{p}) \times w_{d,k+1} \times U(\epsilon, -1, 1) \times \frac{1}{k+1} \times \frac{1}{k} \times \frac{1}{2} \times \left| \frac{\partial(\boldsymbol{\beta}', \epsilon)}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)} \right|$$
$$\times \min \left\{ 1, (k+1) \times \frac{w_{b,k}}{w_{d,k+1}} \times \frac{\pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k)}{\pi(\boldsymbol{\theta}', \mathbf{s}', k+1)} \times 2 \times \left| \frac{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)}{\partial(\boldsymbol{\beta}', \epsilon)} \right| \right\}$$
$$= (1-\tilde{p}) \times U(\epsilon : -1, 1) \times \min \left\{ \pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times w_{d,k+1} \times \frac{1}{k(k+1)} \times \frac{1}{2} \right.$$
$$\left. \times \left| \frac{\partial(\boldsymbol{\beta}', \epsilon)}{\partial(\boldsymbol{\beta}^{(t)}, \epsilon^*, \epsilon)} \right|, \frac{1}{k} \times w_{b,k} \times \pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \right\}$$
$$= (1-\tilde{p}) \times U(\epsilon : -1, 1) \times \min \left\{ \pi(\boldsymbol{\theta}', \mathbf{s}', k+1) \times w_{d,k+1} \times \frac{1}{k(k+1)} \right.$$
$$\left. \times |\beta'_{j'}|, \frac{1}{k} \times w_{b,k} \times \pi(\boldsymbol{\theta}^{(t)}, \mathbf{s}^{(t)}, k) \right\}. \tag{S-4}$$

Noting that $|\beta'_{j'}| = |\beta_j^{(t)}|/|\epsilon|$, it is seen that (S-3) = (S-4); that is, detailed balance holds for the birth and the death moves with respect to the multiplicative transformation. Again, the proof of detailed balance for the no-change move type where the dimension remains unchanged is the same as that of TMCMC.

Also, the proof of detailed balance of the mixing-enhancement step (Step 5. of Algorithm S-2.1) is the same as that of TMCMC.

S-4.2 Irreducibility and aperiodicity

The proof of irreducibility and aperiodicity of Algorithm S-2.1 follows easily from the general arguments provided in the supplements of Das and Bhattacharya (2019) and Dutta and Bhattacharya (2014).

S-5 Bayes factor based variable selection experiments with TTMCMC

We now provide details of our simulation studies with respect to variable selection. We consider linear regression (Section 4.1 of MB), Gaussian process regression with squared exponential covariance kernel (Section 4.2 of MB) as well as autoregressive regression (Section 7.1 of MB) for our purpose.

S-5.1 Linear regression

S-5.1.1 Data generation with random sets of covariates

As in Section 4.1 of MB, we consider the model of the form $y_i = \boldsymbol{\beta}'_{\mathbf{s}}\mathbf{x}_{i,\mathbf{s}} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$. For the true, data-generating model, we set $\sigma^2_\epsilon = 0.1$, and set, for $i = 1, \ldots, n$ and $j = 1, \ldots, p = |\mathbf{S}|$, $x_{ij} = 5/j + \eta_{ij}$, where $\eta_{ij} \overset{iid}{\sim} N(0, \sigma^2_\eta)$, with $\sigma^2_\eta = 0.1$. For generating the data, we randomly select a subset $\mathbf{s}$ from the set $\mathbf{S}$ associated with p covariates, construct $\mathbf{x}_{i,\mathbf{s}}$ and simulate the elements of the regression coefficient vector $\boldsymbol{\beta}_{\mathbf{s}}$ independently from $N(0, \sigma^2_b)$, with $\sigma^2_b = 5$. We also consider an intercept α in our data-generating model, which we simulate as $\alpha \sim N(\mu_\alpha, \sigma^2_\alpha)$, with $\mu_\alpha = 1$ and $\sigma^2_\alpha = 0.1$. Abusing notation for convenience, we shall assume that α is the first element of $\boldsymbol{\beta}_{\mathbf{s}}$ and that the vector of ones is the first column of the design matrix $X_{\mathbf{s}}$. With this setup, we then generate the data from the resulting true regression model.

For data generation, we consider three scenarios. Setting $p = 10, 20, 30$, we generate $n = 25, 25, 35$ data-points for the respective values of p. We repeat the data-generation procedure 1000 times for each pair (p, n), so that for every (p, n), we have 1000 datasets, each consisting of n data-points and a random subset of covariates selected from the possible p covariates. For each of the 1000 simulated datasets, we attempt to select the best subset of covariates using Bayes factor obtained through TTMCMC. The Bayesian model and prior specifications that we used for the purpose is detailed next.

S-5.1.2 Bayesian linear regression model and prior specification for variable selection using TTMCMC and Bayes factors

Then assuming that the model for the simulated data y_i is normal linear regression (with intercept) on an unknown subset of covariates of the complete set of p covariates, and with all parameters unknown, we attempt to select the best subset of covariates, using our TTMCMC algorithm (Algorithm S-2.1) and Bayes factors resulting from TTMCMC, as detailed in Section S-3. For the prior on $\boldsymbol{\beta}_{\mathbf{s}}$, we consider the same form of Zellner's g prior considered in Section 4.1 of MB; here we assume the following equivalent form:

$$\boldsymbol{\beta}_{\mathbf{s}} \sim N\left(\mathbf{0}, \exp(\phi - g)\left(X'_{\mathbf{s}}X_{\mathbf{s}}\right)^{-1}\right), \tag{S-1}$$

where g and ϕ are real-valued parameters. Rather than fixing g and ϕ, we consider them as random variables, to be updated in TTMCMC. Thus, priors are needed on these parameters. As in the case of Zellner-Siow prior

(Zellner and Siow (1980); see also Liang et al. (2008) for further discussion), we assume that *a priori*, $\exp(g) \sim Gamma(1/2, n/2)$, so that the log-prior for g is given, after ignoring an additive constant, by

$$\log \pi(g) = -\frac{n}{2}\exp(g) + \frac{g}{2}. \tag{S-2}$$

We also assume that

$$\pi(\phi) \propto 1. \tag{S-3}$$

As regards the prior for σ_ϵ^2, we re-parameterize this as $\exp(-\tau)$, and assume that $\exp(\tau) \sim Gamma(a_\tau, b_\tau)$, so that the log-prior, after ignoring an additive constant, is given by

$$\log \pi(\tau) = -b_\tau \exp(\tau) + a_\tau \tau. \tag{S-4}$$

We set $a_\tau = b_\tau = 0.01$.

We put a discrete normal prior on $k = |\mathbf{s}|$, given by

$$\pi(k) \propto \exp\left\{-\frac{1}{2\sigma_k^2}(k - \mu_k)^2\right\};\ k = 1, 2, \ldots, p. \tag{S-5}$$

Note that although the Poisson distribution is commonly used for specifying priors on the dimension in variable-dimensional problems, the above discrete normal prior is more flexible, since it can control both the mean and variance of the dimensionality, unlike the Poisson prior which has the same mean and variance.

In (S-5) we set $\mu_k = 8, 16, 24$, respectively, when $p = 10, 20, 30$, and fix $\sigma_k^2 = 1$ for all the chosen values of p. These relatively large values of μ_k with respect to p are chosen to avoid the Lindley's paradox which creates the tendency among Bayes factors to select parsimonious models, irrespective of the truth. The variance $\sigma_k^2 = 1$ is expected to disallow significant drift of the dimension towards small values, unless the data dictates so.

S-5.1.3 TTMCMC implementation for Bayesian linear regression

For TTMCMC implementation, we set $w_{b,k} = w_{d,k} = w_{nc,k} = 1/3$ for all $k = 2, \ldots, p-1$; for $k = 1$ and $k = p$, we set $w_{d,k} = 0$ and $w_{b,k} = 0$, respectively. For the latter two cases, we set $w_{b,k} = w_{nc,k} = 1/2$ and $w_{d,k} = w_{nc,k} = 1/2$, respectively.

We also set $\tilde{p} = \tilde{q} = 1/2$, so that we select additive and multiplicative transformations with equal probabilities. We set the scales $a_{\beta,j} = 0.5$, for $j = 1, \ldots, p$, and $a_{\theta,j} = 0.05$ for the remaining parameters. However, when

p is as large as 20 and 30, we set $a_{\theta,j} = 0.005$ for the remaining parameters to make the acceptance rates reasonably large. For the no-change and mixing-enhancement steps, we set $c = 0.01$. Recall from the discussion in Section S-2 that the goal of this strategy is to improve acceptance rates of the birth moves as well as of the no-change and mixing-enhancing moves, induced by the additive transformation. Indeed, note that with the additive transformation, the acceptance ratio of the birth move depends significantly on twice $a_{\beta,j}$, so that relatively large value of $a_{\beta,j}$ would lead to higher acceptance probability. However, too large $a_{\beta,j}$ would of course lead to increased rejection rate, since $\beta_j^{(t)} + a_{\beta,j}|\epsilon_1|$ and $\beta_j^{(t)} - a_{\beta,j}|\epsilon_1|$ may take the new $\boldsymbol{\beta}$-vector too far from the current $\boldsymbol{\beta}^{(t)}$-vector. Thus, relatively large, but adequate choices of $a_{\beta,j}$s are necessary. This also ensures that the acceptance rate of the death move, which depends upon inverse of $a_{\beta,j}$, is not too small.

Now, relatively large choice of $a_{\beta,j}$s would make the acceptance rates associated with the no-change and the mixing-enhancing steps induced by the additive transformation too small, since in those steps, all the unknown quantities are updated simultaneously. To avoid this undesirable situation, we multiply $a_{\beta,j}$s by $c = 0.01$, so that they are rendered adequately small in these steps. Detailed balance is easily seen to hold with respect to this multiplication by c, in the same way as in fixed-dimensional TMCMC.

We standardize all the available covariates so that their empirical means and variances are 0 and 1, respectively. Now note that minimizing $BIC(u)$ in the linear regression case reduces to minimizing the residual sum of squares $\sum_{i=1}^{n}(y_i - \hat{\boldsymbol{\beta}}'_{\mathbf{s}^{(t)}}\mathbf{x}_{i,\mathbf{s}^{(t)}} - \hat{\beta}_u x_{i,u})^2$ with respect to $u \in \mathbf{S}\backslash\mathbf{s}^{(t)}$, where $(\hat{\boldsymbol{\beta}}_{\mathbf{s}^{(t)}}, \hat{\beta}_u)$ is the least squares estimator associated with the current covariate index subset $\mathbf{s}^{(t)}$.

For our TTMCMC implementation, we discard the first $10^4 \times 150$ iterations as burn-in, and store one in every 150 iterations in the next $5\times10^4\times150$ iterations, to obtain 5×10^4 iterations for our Bayesian inference. We initialise our TTMCMC algorithm with only one covariate, $\{x_{i1};\ \ = 1, \ldots, n\}$.

S-5.1.4 Parallelization

Recall that for every pair (p, n), 1000 datasets are generated and TTMCMC must be implemented for variable selection in each of the 1000 datasets. Thus, 1000 TTMCMC implementations are necessary for each pair (p, n). For Bayesian linear regression, a single typical TTMCMC run in our C code implementation on each core (with 2.8 GHz CPU speed) of our VMWare

(about 2 TB memory) takes about 2 minutes, 4 minutes and 11 minutes, respectively, for $(p = 10, n = 25)$, $(p = 20, n = 25)$ and $(p = 30, n = 35)$. Hence, for completing our simulation experiments in reasonable times, parallelization of our computations is indispensable.

Although our VMWare that we use for our current research consists of 80 single-threaded cores, using only the best 50 of them yields the optimum performance. As such, using shell scripting language, we parallelise the 1000 C code based TTMCMC runs for each (p, n) combination into 50 cores, so that 50 TTMCMC runs are simultaneously implemented for each (p, n); each core implementing only 20 TTMCMC runs. This parallelization strategy allowed us to obtain the results for all our simulation experiments in very reasonable times, as is obvious from the aforementioned timings for the single TTMCMC runs.

A typical TTMCMC run for the $(p = 10, n = 25)$ case yields the overall acceptance rate 0.193, birth rate 0.064, death rate 0.081 and no-change rate 0.410. For $(p = 20, n = 25)$, these rates are 0.191, 0.039, 0.039 and 0.496, respectively, and for $(p = 30, n = 35)$, these are 0.234, 0.101, 0.101 and 0.498, respectively. These rates are computed on the basis of the entire TTMCMC run, not just on the stored samples. That is, these rate computations are based on $10^4 \times 150 + 5 \times 10^4 \times 150 = 9 \times 10^6$ TTMCMC realizations.

S-5.1.5 Results of the linear regression simulation experiments

After every TTMCMC run in each processor of our VMWare, we implement an R code that computes B_i given by (S-2), for $i = 1, \ldots \tilde{N}$. The R code selects that set of covariates indexed by $\mathbf{s}_{best}$ which corresponds to $B_{\max} = \max\{B_1, \ldots, B_{\tilde{N}}\}$. We also consider a binary vector $V = (v_1, \ldots, v_p)$, where, for $j = 1, \ldots, p$, $v_j = 1$ or 0 accordingly as $j \in \mathbf{s}_{best}$ or $j \notin \mathbf{s}_{best}$. Also, let V_0 denote the binary vector associated with $\mathbf{s}_0$, the set of indices of the data-generating covariates. The R code also computes the Hamming distance between the binary vectors V and V_0, which, simply put, is the total number of position-wise mismatches in the two vectors consisting of p positions. Thus, the Hamming distance is zero if and only if $V = V_0$, that is, when the best model obtained is the same as the true model. The Bayes factor of the best model against the true, data-generating model is also computed in the R code using the formula (S-3), provided that the true model appears in the stored TTMCMC sample. Furthermore, we also compute the rank of the true model based on the B_i values, again provided that the true model features in the stored TTMCMC sample.

For each (p, n), these results for all the 1000 TTMCMC runs are com-

bined to yield the proportions of times the Hamming distance takes the values $0, 1, \ldots, p-1$, among the 1000 runs. We also compute the average log-Bayes factor of the best model against the true model and the average rank of the true model, the averaging done over those TTMCMC samples which consist of the true model and Hamming distance value r, for $r = 0, 1, \ldots, p-1$. These results are depicted in Figure S-1. Panel (a) of the figure shows that for $p = 10, n = 25$, the Hamming distance gives the highest probability (about 0.527) to 0, that is, the true set of covariates is selected with the highest probability, which is also significantly higher compared to the other values of the Hamming distance. The average log-Bayes factor, as shown in panel (b), is the highest when the Hamming distance is 5, while for Hamming distance 7, 8 and 9, the average log-Bayes factor is not available since the true set of covariates did not appear in the TTMCMC samples in those cases. Note that the average log-Bayes factor is not increasing with the Hamming distance, which is indeed not to be expected in general. The average true model rank, displayed in panel (c), is increasing with the Hamming distance, but again, is unavailable for the values 7, 8 and 9 of the Hamming distance since the true set of covariates has probability zero with respect to the respective TTMCMC samples.

The scenario when $p = 20, n = 25$, is not significantly different from the $p = 10, n = 25$ case. Panel (d) shows that the Hamming distance gives the highest probability 0.143 to both 0 and 1, which is again significantly higher than those for the other values. The highest probability is of course much less than in the corresponding $(p = 10, n = 25)$ scenario, which is expected, since the number of covariate subsets to search for the true covariate subset is far greater than in the previous case. Since $n = 25$ is also the same as before, the information about the true covariate set is not increased either. But that in spite of these issues the true covariate sets are found with the highest probability, vindicates the efficacy of our variable selection theory and the TTMCMC based methodology. The average log-Bayes factor is the highest when the Hamming distance is 9 and the average true model rank is the highest for Hamming distance 7. Note that unlike the case of $(p = 10, n = 25)$, the true model rank is not increasing with the Hamming distance in this case, and this is to be expected in general

The case of $p = 30, n = 35$ is the most challenging situation among all the (p, n) pairs considered, as searching for the true set of covariates from among a set of $2^{30} - 1 = 1073741823$ possible subsets is akin to looking for a needle in a haystack! Yet, as panel (g) of Figure S-1 shows, the Hamming distance gives significant probability to 0, while the value 2 gets the highest probability. This performance of our TTMCMC based Bayes factor should

not be considered unsatisfactory at all. Note that the average log-Bayes factor is the highest when the Hamming distance is 10, but the average model rank is the worst when the Hamming distance is as small as 5, relative to $p-1=29$. In other words, even for Hamming distance 5, there are many models that perform better than the true model on a average, in terms of Bayes factor. Since for most of the larger values of the Hamming distance the true sets of covariates have probabilities zero with respect to TTMCMC, it is clear that for most values of the Hamming distance, the true, data-generating model is outperformed by the other models. It must also be remarked that for a limited TTMCMC sample size, reliably measuring the performances of many important models among a set of such a huge number of models, is infeasible.

S-5.2 Gaussian process regression

We now consider simulation experiments with Gaussian process regression as described in Section 4.2 of MB. That is, now the model that we consider is of the form $y = f(\mathbf{x_s}) + \epsilon$, where $f(\cdot)$ is modeled by a Gaussian process with mean function $\mu(\mathbf{x_s})$ and squared exponential covariance kernel of the form

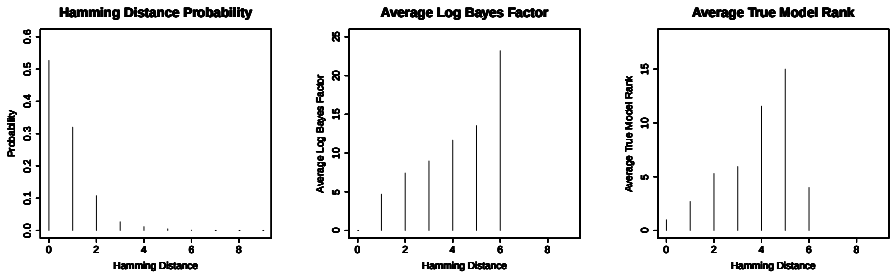
$$Cov\left(f(\mathbf{x_s}), f(\mathbf{x'_s})\right) = \sigma_f^2 \exp\left\{-\frac{1}{2}\left(\mathbf{x_s}-\mathbf{x'_s}\right)^T D_{\mathbf{s}}\left(\mathbf{x_s}-\mathbf{x'_s}\right)\right\},$$

where σ_f^2 is the process variance and the diagonal elements of $D_{\mathbf{s}}$ are the smoothness parameters.

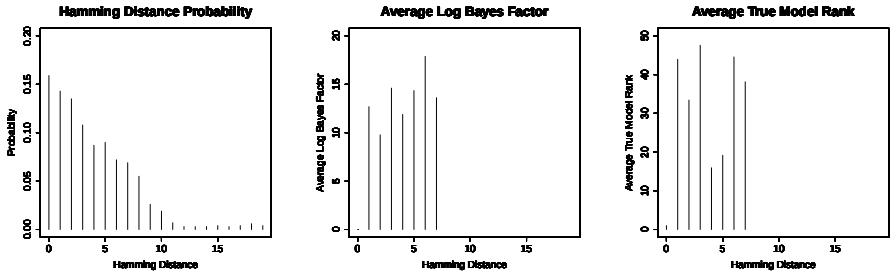
Here the data are modeled as $y_i = f(\mathbf{x}_{i,\mathbf{s}}) + \epsilon_i$, where, for $i = 1, \ldots, n$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma_\epsilon^2)$. As before, we reparameterize σ_ϵ^2 as $\exp(-\tau)$; we also reparameterize σ_f^2 as $\exp(-\tau_f)$, where τ and τ_f are real parameters. But unlike the linear regression case, here we assume that *a priori*, $\tau \sim N(\mu_\tau, \sigma_\tau^2)$ and $\tau_f \sim N(\mu_{\tau_f}, \sigma_{\tau_f}^2)$, with $\mu_\tau = \mu_{\tau_f} = 0$, $\sigma_\tau^2 = 0.5$ and $\sigma_{\tau_f}^2 = 0.1$, the variances reflecting the belief that uncertainty about the process variance is less than that of the noise variance.

For $i = 1, \ldots, |\mathbf{s}|$, we reparameterize the i-th diagonal element of $D_{\mathbf{s}}$ as $\exp(-\gamma_i)$, where we assume *a priori* that $\gamma_i \overset{iid}{\sim} N(0, \sigma_\gamma^2)$, with $\sigma_\gamma^2 = 2$.

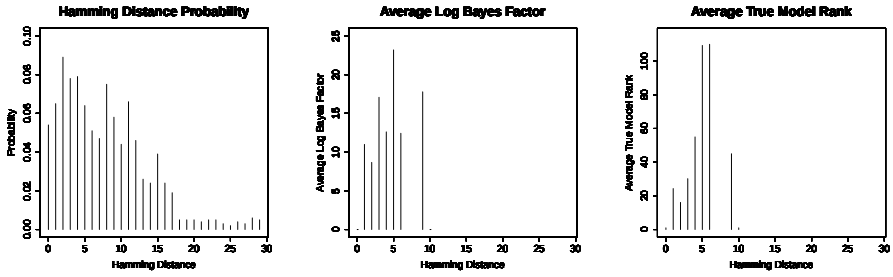
We model the mean function $\mu(\mathbf{x_s})$ as linear regression containing the intercept, that is, we set $\mu(\mathbf{x_s}) = \boldsymbol{\beta}'_{\mathbf{s}}\mathbf{x_s}$, assuming that the first element of $\mathbf{x_s}$ is 1. We consider the same Zellner-Siow prior form for $\boldsymbol{\beta}_{\mathbf{s}}$ as in Section S-5.1. As before, we standardize all the available covariates for model implementation with TTMCMC.



(a) $p = 10, n = 25$. (b) $p = 10, n = 25$. (c) $p = 10, n = 25$.



(d) $p = 20, n = 25$. (e) $p = 20, n = 25$. (f) $p = 20, n = 25$.



(g) $p = 30, n = 35$. (h) $p = 30, n = 35$. (i) $p = 30, n = 35$.

Figure S-1: Simulation study: Bayesian linear regression variable selection results.

Letting $\boldsymbol{\gamma}_{\mathbf{s}}$ denote the vector of smoothness parameters, note that $\boldsymbol{\beta}_{\mathbf{s}}$ and $\boldsymbol{\gamma}_{\mathbf{s}}$ are both variable-dimensional vectors, the dimensions of which must be increased or decreased simultaneously. Recall that such updating provision is of course considered in our TTMCMC algorithm (Algorithm S-2.1). The prior for k remains the same as in the linear regression setup.

The data simulation principle from the true model consisting of random sets of covariates and the formation of the covariates remain the same as in the linear regression case; here $(y_1, \ldots, y_n)$ is generated from the joint multivariate normal model dictated by the above Gaussian process setup, given $\boldsymbol{\beta}_{\mathbf{s}}$, $\boldsymbol{\gamma}_{\mathbf{s}}$, τ and τ_f. We set $\tau = -\log(0.1)$ and $\tau_f = -\log(0.2)$, and simulate the elements of $\boldsymbol{\beta}_{\mathbf{s}}$ and $\boldsymbol{\gamma}_{\mathbf{s}}$ from the zero mean normal distribution with variance 5.

As before, we consider the settings $(p = 10, n = 25)$, $(p = 20, n = 25)$ and $(p = 30, n = 35)$ for evaluating our Bayes factor based variable selection obtained via TTMCMC. The TTMCMC algorithm in this Gaussian process setup is similar to that for linear regression, with the extra variable-dimensional parameter $\boldsymbol{\gamma}_{\mathbf{s}}$ and the fixed-dimensional variable τ_f being accounted for. The procedure for updating these remain the same as before, in accordance with the details provided in Algorithm S-2.1.

As regards computation of $BIC(u)$ in this Gaussian process setup, we first obtain the least squares estimates $\hat{\boldsymbol{\beta}}_{\mathbf{s}}$ corresponding to $\boldsymbol{\beta}_{\mathbf{s}}$ pretending a linear regression context, and substitute $\hat{\boldsymbol{\beta}}_{\mathbf{s}}$ in the Gaussian process likelihood. Also, in the Gaussian process likelihood, we set $\hat{\gamma}_i = 0$, for $i = 1, \ldots, |\mathbf{s}|$, corresponding to $\boldsymbol{\gamma}_{\mathbf{s}}$. Finally, we substitute the current values $\tau^{(t)}$ and $\tau_f^{(t)}$ for τ and τ_f in the likelihood, and proceed to compute the BIC version with these substitutions. As we shall demonstrate, our experiments reveal that this method yields quite reliable propositions for the new covariates. However, computation of $BIC(u)$ is quite demanding in this setup due to the requirement of $n \times n$-order matrix inversions for every u. Thus, in this setup, for p even moderately large, our TTMCMC takes considerably more implementation time than for linear regression. Indeed, for $(p = 10, n = 25)$ the time taken is about 26 minutes for a typical run, and for $(p = 20, n = 25)$ and $(p = 30, n = 35)$, the respective run times are about 46 minutes and 2 hours 22 minutes. We parallelise our simulation experiments consisting of 1000 TTMCMC runs for each (p, n) pair in the same way as in the linear regression setup.

The overall acceptance rate, birth rate, death rate and the no-change rate for a typical TTMCMC run in the $(p = 10, n = 25)$ case are about 0.182, 0.014, 0.014 and 0.518, respectively. For $(p = 20, n = 25)$, these

numbers are 0.217, 0.065, 0.065 and 0.522, while in the $(p = 30, n = 35)$ scenario, the respective rates are 0.201, 0.046, 0.046 and 0.511. Again, these rates are computed on the basis of 9×10^6 TTMCMC realizations.

S-5.2.1 Results of the Gaussian process regression simulation experiments

The results of our variable selection method in the Gaussian process setup are encapsulated in Figure S-2. Note that although for $(p = 10, n = 25)$ the Hamming distance assigns significantly higher probability to 0 compared to all the other values, in the other two more challenging scenarios $(p = 20, n = 25)$ and $(p = 30, n = 35)$, this good performance is not kept up. This observation is in line with the average log-Bayes factors and the average true model ranks, as displayed in Figure S-2. Thus, compared to the linear regression setup, our variable selection methods in the Gaussian process regression setup seems to be less robust with respect to increasing dimensions. This is, however, not unexpected due to the structured dependence in the Gaussian process regression datasets.

S-5.3 AR(1) regression

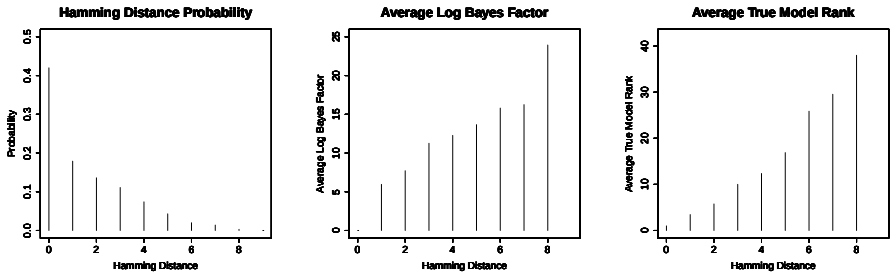
Now let us consider the variable selection problem in the following $AR(1)$ context, the Bayes factor asymptotics of which is detailed in Section 7.1:

$$y_t = \rho y_{t-1} + \boldsymbol{\beta}_{\mathbf{s}}' \mathbf{x}_{t,\mathbf{s}} + \epsilon_t, \quad \text{and} \quad \epsilon_t \overset{iid}{\sim} N\left(0, \sigma_\epsilon^2\right), \quad \text{for} \quad t = 1, \ldots, n,$$

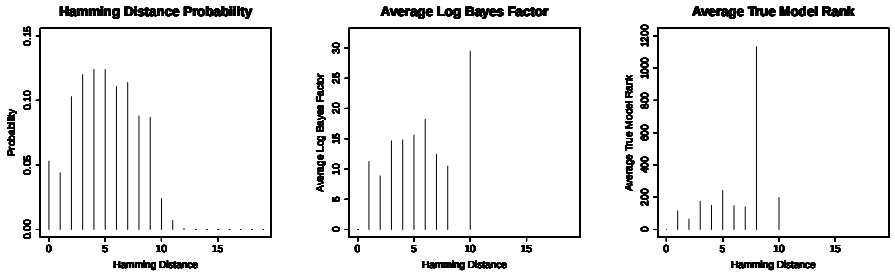
where $y_0 \equiv 0$ and $|\rho| < 1$. We reparameterize ρ as $\rho = -1 + 2\exp(\tilde{\rho})/(1 + \exp(\tilde{\rho}))$, and assume that $\tilde{\rho} \sim N(0, 1)$. As before we consider the Zellner-Siow prior for $\boldsymbol{\beta}_{\mathbf{s}}$ and reparameterize σ_ϵ^2 as $\exp(-\tau)$ and use the log-prior form (S-4), with $\alpha_\tau = \lambda_\tau = 0.01$. For data generation, we generate ρ from $U(-1, 1)$.

To form the likelihood, we considered the product $\prod_{t=1}^{n} [y_t | y_{t-1}]$, where $[y_t | y_{t-1}]$ stands for the distribution (19) of MB. We did not consider the correlated error form $y_t = \boldsymbol{\beta}_{\mathbf{s}}' \boldsymbol{z}_{t,\mathbf{s}} + \tilde{\epsilon}_t$ considered in Section 7.1 for likelihood formation as this would involve a multivariate normal distribution which would require $n \times n$ matrix inversions in each step of TTMCMC.

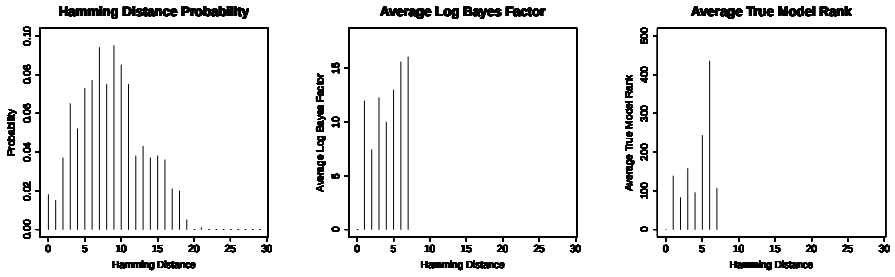
For $BIC(u)$, first note that given the value of ρ at the current TTMCMC iteration, y_1 and $y_t - \rho y_{t-1}$ for $t \geq 2$ has a linear regression form, using which we compute the least squares estimator of the regression coefficients. The rest of the $BIC(u)$ minimization procedure is the same as in the linear regression setup.



(a) $p = 10, n = 25$. (b) $p = 10, n = 25$. (c) $p = 10, n = 25$.



(d) $p = 20, n = 25$. (e) $p = 20, n = 25$. (f) $p = 20, n = 25$.



(g) $p = 30, n = 35$. (h) $p = 30, n = 35$. (i) $p = 30, n = 35$.

Figure S-2: Simulation study: Bayesian Gaussian process regression variable selection results.

Indeed, the remaining methodological and implementation details are also akin to the linear regression situation. In this case, typical TTMCMC runs for $(p = 10, n = 25)$, $(p = 20, n = 25)$ and $(p = 30, n = 35)$ took about 9 minutes, 14 minutes and 36 minutes, respectively. The overall acceptance rate, birth rate, death rate and no-change rates in the first case was about 0.553, 0.552, 0.564 and 0.544, respectively. In the second and third scenarios they were $(0.582, 0.602, 0.601, 0.543)$ and $(0.580, 0.605, 0.605, 0.533)$, respectively.

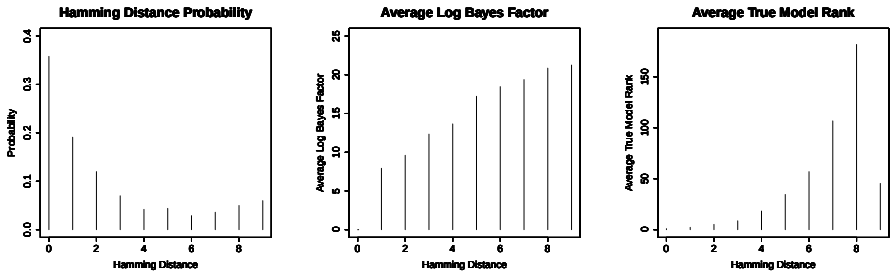
S-5.3.1 Results of the AR(1) regression simulation experiments

Figure S-3 displays the results of our variable selection experiments in the AR(1) context. Note that for $(p = 10, n = 25)$, the Hamming distance gives the highest probability to 0, which is also significantly higher than those for the other values. However, as in the Gaussian process regression experiments, here also the situation deteriorates for $(p = 20, n = 25)$ and $(p = 30, n = 35)$, as the Hamming distance concentrates around larger and larger values for the latter two scenarios. This gradual worsening of the performance is also reflected in the respective average log-Bayes factors and the average true model ranks, demonstrating that compared to linear regression, variable selection in time series regression is a much more delicate problem, with marked sensitivity with respect to larger dimensions.

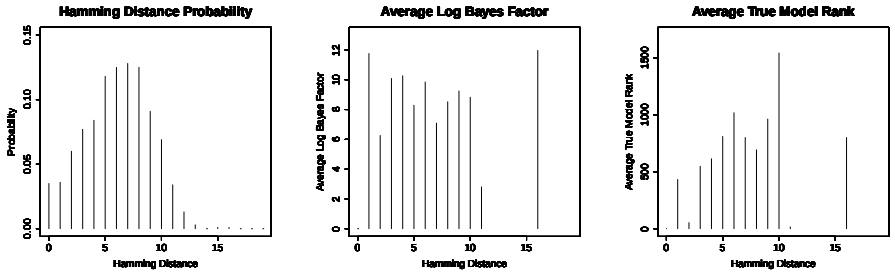
At the first glance, this lack of robustness with respect to dimension might seem surprising since the structure of AR(1) regression closely resembles that of linear regression. However, recall from Section 7.1 of MB that although the data y_t can be written in a linear regression form with modified covariate structures involving ρ, the regression errors in such as case are correlated, rendering the AR(1) setup a Gaussian process structure. Hence, it is not surprising that our AR(1) regression results are much more in resemblance with our Gaussian process regression results, as compared to the linear regression results, when $p = 20$ and 30.

S-6 Variable selection in a real riboflavin dataset

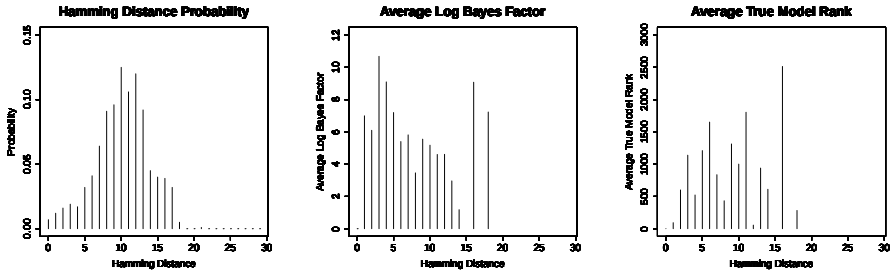
Let us now apply our TTMCMC based variable selection procedure to a real dataset on riboflavin (vitamin B_2) production rate, which has been made public by Bühlmann et al. (2014). The response variable in this dataset is the log-transformed riboflavin production rate corresponding to $p = 4088$ possible covariates measuring the logarithm of the expression level of 4088 genes. The sample size is $n = 71$. Covariate selection of this dataset using



(a) $p = 10, n = 25$. (b) $p = 10, n = 25$. (c) $p = 10, n = 25$.



(d) $p = 20, n = 25$. (e) $p = 20, n = 25$. (f) $p = 20, n = 25$.



(g) $p = 30, n = 35$. (h) $p = 30, n = 35$. (i) $p = 30, n = 35$.

Figure S-3: Simulation study: Bayesian AR(1) regression variable selection results.

classical methods based on linear regression model has been performed by Javanmard and Montanari (2014) who report two significant genes YXLD_at and YXLE_at. Meinshausen et al. (2009) on the other hand, found only YXLD_at to be significant, while the method of Bühlmann (2013) found no significant gene. Thus, based on linear regression models and classical methods of variable selection employed so far, either no gene, one gene or two genes, are found to be significant.

We apply our Bayes factor based covariate selection technique using TTMCMC to this dataset, considering both linear and Gaussian process regression. We consider the same setups as in our simulation experiments, with the same models and priors, with some variation in the prior for k, to account for the uncertainty with respect to the large number of available covariates. Specifically, with the same discrete normal prior for k, we consider the choices of (μ_k, σ_k^2) to be $(25, 10)$ and $(40, 10)$, respectively. We allow a maximum of 50 covariates in the model, since, as is clear from the aforementioned past analyses of this dataset, too many genes can not be significant. In keeping with this, the maximum number of covariates in our posterior simulations turned out to be less than 10 in all our setups for this real data.

As in the simulation experiments, we discard the first $10^4 \times 150$ TTMCMC realizations as burn-in and store every 150-th realization in the next $5 \times 10^4 \times 150$ iterations, to obtain 5×10^4 TTMCMC realizations for our inference. For the additive transformation, the scales $a_{\theta,j}$ and the constant c in the mixing-enhancing step are chosen in the same way as in the simulation studies in the $p = 30$ setups. Since for Gaussian process regression minimization of $BIC(u)$ in the TTMCMC step is computationally too demanding for $p = 4088$ covariates, we replace this $BIC(u)$ with that used for linear regression. Theoretically, this is a perfectly valid procedure, and our TTMCMC results demonstrate very reasonable final selection of the covariates via Bayes factor.

S-6.1 Results for the linear regression model

For the linear regression model with $(\mu_k, \sigma_k^2) = (25, 10)$, the overall acceptance rate, birth rate, death rate, no-change rate turned out to be $(0.187, 0.002, 0.003, 0.497)$, and the implementation time is about 12 hours and 11 minutes. Our Bayes factor computation based on TTMCMC yielded the following best set of 4 covariates: (ARGB_at, EXOA_at, SIGY_at, YOAB_at).

When $(\mu_k, \sigma_k^2) = (40, 10)$, the overall acceptance rate, birth rate, death rate, no-change rate are $(0.175, 0.001, 0.001, 0.496)$ and the time taken is

about 11 hours 33 minutes. In this case, the following set of 7 covariates turned out to be the best: (ARGB_at, YDAR_at, YHDZ_at, YJIA_at, YOAB_at, YUZF_at, YXLD_at). The covariates common to both $(\mu_k, \sigma_k^2) = (25, 10)$ and $(\mu_k, \sigma_k^2) = (40, 10)$ are (ARGB_at,YOAB_at), which does not contain the covariates found significant by Javanmard and Montanari (2014) or Meinshausen et al. (2009). More specifically, although the case $(\mu_k, \sigma_k^2) = (40, 10)$ contains YXLD_at, which has been found to be significant by both Javanmard and Montanari (2014) and Meinshausen et al. (2009), none of the cases $(\mu_k, \sigma_k^2) = (25, 10)$ or $(\mu_k, \sigma_k^2) = (40, 10)$ finds YXLE_at, declared as significant by Javanmard and Montanari (2014).

S-6.2 Results for the Gaussian process regression model

For the Gaussian process regression model with $(\mu_k, \sigma_k^2) = (25, 10)$, the implementation time is about 8 hours 47 minutes and the overall acceptance rate, birth rate, death rate, no-change rate are $(0.185, 0.012, 0.012, 0.530)$. The following 4 covariates are selected as the best by our TTMCMC based Bayes factor: (ARGB_at, YHDZ_at, YOAB_at, YXLD_at).

For $(\mu_k, \sigma_k^2) = (40, 10)$, the implementation time was about 9 hours 56 minutes and the overall acceptance rate, birth rate, death rate, no-change rate are $(0.193, 0.024, 0.024, 0.529)$. Remarkably, here we obtain exactly the same set of covariates (ARGB_at, YHDZ_at, YOAB_at, YXLD_at), as for $(\mu_k, \sigma_k^2) = (25, 10)$, as the best set of covariates, which exhibits considerable robustness of the Gaussian process regression model with respect

S-6.3 Comparison of the results for the linear regression and the Gaussian process regression models

The results demonstrate that compared to the linear regression model for this data, the Gaussian process regression is far more robust with respect to the prior for k; moreover, it leads to much parsimony compared to linear regression, as can be easily seen from the cardinalities of the best covariate sets.

Note that YOAB_at is common to all our linear regression and Gaussian process regression implementations and hence we consider this to be an important discovery. Also, YXLD_at is common to our Gaussian process and linear regression implementations with $(\mu_k, \sigma_k^2) = (40, 10)$. Since this gene is found to be significant by Javanmard and Montanari (2014) and Meinshausen et al. (2009) as well, it seems that this may also be an important discovery. But the YXLE_at gene, although declared significant by Javan-

mard and Montanari (2014), did not appear in the best set of covariates in any of our linear regression or Gaussian process regression analysis.

Now, the question arises that which of the four implementations of linear and Gaussian process regression yields the best result in terms of Bayes factor. In this regard, we first note that the maximum values of the log of B_i given by (S-2) associated with the linear regression models for $(\mu_k, \sigma_k^2) = (25, 10)$ and $(\mu_k, \sigma_k^2) = (40, 10)$ are given by 51.925 and 94.697, respectively, while the same for the Gaussian process regression counterpart are 46.085 and 91.147. Thus, in this regard, the linear regression model with $(\mu_k, \sigma_k^2) = (40, 10)$ given its Bayes factor guided best possible set of covariates, is the best model, followed by Gaussian process regression with $(\mu_k, \sigma_k^2) = (40, 10)$, given its best set of covariates. The next best models, given their respective best set of covariates with respect to Bayes factors, in order, are provided by the linear regressions model with $(\mu_k, \sigma_k^2) = (25, 10)$ and the Gaussian process regression model with $(\mu_k, \sigma_k^2) = (25, 10)$.

Note that the genes YOAB_at and YXLD_at are common to the best two models, which once again vindicates their importance.

S-7 Appendix

In this section we provide the proofs of all the lemmas stated in the paper. Before proving the lemmas, we state some results which are useful in proving the lemmas.

Result S-1. *Let $\mathbb{C}^{m\times n}$ denote the vector space of all $m \times n$ matrices. If $G, H \in \mathbb{C}^{n\times n}$ are positive semidefinite Hermitian matrices and $1 \leq i_1 < \cdots < i_k \leq n$, then the following two inequalities hold:*

$$\sum_{t=1}^{k} \lambda_{i_t}(GH) \leq \sum_{t=1}^{k} \lambda_{i_t}(G)\,\lambda_t(H)$$
$$\sum_{t=1}^{k} \lambda_t(GH) \geq \sum_{t=1}^{k} \lambda_{i_t}(G)\,\lambda_{n-i_t+1}(H).$$

A proof of this result can be found in Wang and Zhang (1992).

Result S-2. *(a) For matrices A_1 and A_2, let $A_1 \preceq A_2$ imply that $A_2 - A_1$ is nonnegative definite. Then for any symmetric matrix A,*

$$\lambda_{\min}(A)I \preceq A \preceq \lambda_{\max}(A)I,$$

(b) For symmetric matrices A_1 and A_2,

$$\lambda_{\min}(A_1)+\lambda_{\min}(A_2) \leq \lambda_{\min}(A_1+A_2) \leq \lambda_{\max}(A_1+A_2) \leq \lambda_{\max}(A_1)+\lambda_{\max}(A_2).$$

Lemma S-1. *Consider the setup of Section 7.1. The eigenvalues $Cov\,(\tilde{\epsilon}_{t+h}, \tilde{\epsilon}_t) = \sigma_\epsilon^2(1-\rho^2)^{-1}\Sigma_\epsilon$ are all positive, and bounded.*

Proof. Let $Cov\,(\tilde{\epsilon}_{t+h}, \tilde{\epsilon}_t) = \sigma_\epsilon^2(1-\rho^2)^{-1}\Sigma_\epsilon$. We first find the highest and lowest eigenvalues of Σ_ϵ. It can be shown that the inverse of Σ_ϵ is a tridiagonal matrix as follows:

$$\left(1-\rho^2\right)\Sigma_\epsilon^{-1} = \begin{pmatrix} 1 & -\rho & 0 & 0 & \dots & 0 & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 \\ & \dots & & \dots & & \dots & \\ 0 & 0 & 0 & 0 & \dots & 1+\rho^2 & -\rho \\ 0 & 0 & 0 & 0 & \dots & -\rho & 1 \end{pmatrix}$$

Stroeker (1983) shows that the approximations of the eigenvalues of $\left(1-\rho^2\right)\Sigma_\epsilon^{-1}$ (arranged in increasing order if $\rho > 0$ and in decreasing order if $\rho < 0$) are

$$\lambda_k\left(\left(1-\rho^2\right)\Sigma_\epsilon^{-1}\right) \approx 1 - 2\rho\cos\frac{k\pi}{n+1} + \rho^2 - \frac{4}{n+1}\rho^2\left(\sin\frac{k\pi}{n+1}\right)^2,$$

with corresponding error bound

$$\xi_k = \frac{2\rho^2}{\sqrt{n+1}}\sin\frac{k\pi}{n+1}, \quad \text{for } k = 1, \dots, n.$$

Combining the above facts, it can be seen that the eigenvalues of $\left(1-\rho^2\right)\Sigma_\epsilon^{-1}$ are bounded by $\min\{(1+\rho)^2, (1-\rho)^2\}+o(1)$ and $\max\{(1+\rho)^2, (1-\rho)^2\}+o(1)$. Thus the eigenvalues of $\left(1-\rho^2\right)^{-1}\Sigma_\epsilon$ are bounded by $\left\{(1-|\rho|)^2+o(1)\right\}^{-1}$ and $\left\{(1+|\rho|)^2+o(1)\right\}^{-1}$. Thus all the eigenvalues of Σ_ϵ are finite. As $|\rho| < 1-\gamma$ for some small enough γ, clearly for sufficiently large n all the eigenvalues of $\left(1-\rho^2\right)\Sigma_\epsilon^{-1}$ are positive. Also the highest eigenvalue is less than $2/\gamma^2$. □

Lemma S-2. *Consider the setup in Section 7.1. Let $P_{n,\mathbf{s}}$ be the orthogonal projection matrix onto the column space of $Z_{n,\mathbf{s}}$. Then the eigenvalues of $\partial P_{n,\mathbf{s}}/\partial\rho$ are uniformly bounded.*

Proof. For simplicity we write Z instead of $Z_{n,\mathbf{s}}$. Recall $P_{n,\mathbf{s}} = Z\left(Z^TZ\right)^{-1}Z^T$. Thus,

$$\begin{aligned}\frac{\partial P_{n,\mathbf{s}}}{\partial \rho} &= \frac{\partial Z}{\partial \rho}\left(Z^TZ\right)^{-1}Z^T - Z\left(Z^TZ\right)^{-1}\frac{\partial\left(Z^TZ\right)}{\partial\rho}\left(Z^TZ\right)^{-1}Z^T \\ &\quad + Z\left(Z^TZ\right)^{-1}\frac{\partial Z^T}{\partial\rho} \qquad \text{(S-1)}\end{aligned}$$

First note that the t-th row of Z, $\mathbf{z}_{t,\mathbf{s}} = \sum_{k=1}^{t}\rho^{t-k}\mathbf{x}_{k,\mathbf{s}}$. Thus,

$$Z^TZ = n\sum_{t=1}^{n}\sum_{k_1=1}^{t}\sum_{k_2=1}^{t}\rho^{2t-k_1-k_2}\left(\frac{\mathbf{x}_{k_1,\mathbf{s}}\mathbf{x}_{k_2,\mathbf{s}}^T}{n}\right). \qquad \text{(S-2)}$$

Since the covariates lie in a compact space, the elements of the $|\mathbf{s}| \times |\mathbf{s}|$ matrices $\mathbf{x}_{k_1,\mathbf{s}}\mathbf{x}_{k_2,\mathbf{s}}^T$ in (S-2) are uniformly bounded. Further note that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\sum_{k_1=1}^{t}\sum_{k_2=1}^{t}\rho^{2t-k_1-k_2} &\leq \frac{1}{n}\sum_{t=1}^{n}\left(\sum_{k_1=1}^{t}|\rho|^{t-k}\right)^2 = \frac{1}{n(1-|\rho|)^2}\sum_{t=1}^{n}\left(1-|\rho|^t\right)^2 \\ &= (1-|\rho|)^{-2}\left[1 - 2\frac{|\rho|}{n}\frac{(1-|\rho|^n)}{(1-|\rho|)} + \frac{\rho^2\left(1-\rho^{2n}\right)}{n\left(1-|\rho|^2\right)}\right].\end{aligned}$$

As $\rho \in [-1+\gamma, 1-\gamma]$, the above facts imply $Z^TZ = nB_n$, where B_n is a $|\mathbf{s}| \times |\mathbf{s}|$ matrix whose elements are uniformly bounded. It follows that $\left(Z^TZ\right)^{-1} = n^{-1}B_n^{-1} = n^{-1}C_{n,\mathbf{s}}$, where the elements of $C_{n,\mathbf{s}}$ are uniformly bounded.

Let $\partial Z/\partial\rho = H_n$. Then the t-th row of H_n is $\mathbf{h}_{t,\mathbf{s}} = \sum_{k=1}^{t}(t-k)\rho^{t-k-1}\mathbf{x}_{k,\mathbf{s}}$. then the first term of (S-1),

$$D := \frac{\partial Z}{\partial\rho}\left(Z^TZ\right)^{-1}Z^T = n^{-1}\begin{pmatrix}\mathbf{h}_{1,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{z}_{1,\mathbf{s}} & \dots & \mathbf{h}_{1,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{z}_{n,\mathbf{s}} \\ \dots & & \dots \\ \mathbf{h}_{n,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{z}_{1,\mathbf{s}} & \dots & \mathbf{h}_{n,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{z}_{n,\mathbf{s}}\end{pmatrix}.$$

The (t_1,t_2)-th element of D is

$$\begin{aligned}d_{t_1,t_2} &= \frac{1}{n}\mathbf{h}_{t_1,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{z}_{t_2,\mathbf{s}} = \frac{1}{n}\left[\sum_{k=1}^{t_1}(t_1-k)\rho^{t_1-k-1}\mathbf{x}_{k,\mathbf{s}}^T\right]C_{n,\mathbf{s}}\left[\sum_{k=1}^{t}\rho^{t-k}\mathbf{x}_{k,\mathbf{s}}\right] \\ &= \frac{1}{n}\sum_{k_1=1}^{t_1}\sum_{k_1=1}^{t_2}(t_1-k_1)\rho^{t_1-k_1-1+t_2-k_2}\mathbf{x}_{k_1,\mathbf{s}}^TC_{n,\mathbf{s}}\mathbf{x}_{k_2,\mathbf{s}}.\end{aligned}$$

Next note that $\mathbf{x}_{k_1,\mathbf{s}}^T C_{n,\mathbf{s}} \mathbf{x}_{k_2,\mathbf{s}} = \sum_{j=1}^{|\mathbf{s}|} \sum_{l=1}^{|\mathbf{s}|} C_{j,l} x_{k_1,j} x_{k_2,l}$. As elements of $C_{n,\mathbf{s}}$, as well as, the covariate space are uniformly bounded, the above quadratic is bounded for any (k_1, k_2).
Now,

$$\begin{aligned}
\frac{1}{n} & \sum_{k_1=1}^{t_1} \sum_{k_1=1}^{t_2} (t_1 - k_1) \rho^{t_1 - k_1 - 1 + t_2 - k_2} \\
\leq & \frac{1}{n} \left(\sum_{k_1=1}^{t_1} (t_1 - k_1) |\rho|^{t_1 - k_1 - 1} \right) \left(\sum_{k_1=1}^{t_2} \rho^{t_2 - k_2} \right) \\
= & \frac{1}{n} \left\{ \frac{1 - |\rho|^{t_1} \left(t_1 + |\rho| - t_1 |\rho| \right)}{(1 - |\rho|)^2} \right\} \left(\frac{1 - |\rho|^{t_2}}{1 - |\rho|} \right). \qquad \text{(S-3)}
\end{aligned}$$

By Gerschgorin circle theorem the maximum eigenvalue of D is bounded by

$$\max_{t_1 \in \{1,\ldots,n\}} \frac{1}{n} \left\{ \frac{1 - |\rho|^{t_1} \left(t + |\rho| - t|\rho| \right)}{(1 - |\rho|)^2} \right\} \sum_{t_2=1}^{n} \left(\frac{1 - |\rho|^{t_2}}{1 - |\rho|} \right) \leq (1 - |\rho|)^{-3},$$

and the minimum eigenvalue is bounded below by $-(1 - |\rho|)^{-3}$. As $\rho \in [-1 + \gamma, 1 - \gamma]$, eigenvalues of D are uniformly bounded.

Next note that A and A^T have same eigenvalues. As the third term of (S-1) is the transpose of the first term, the third term also has finite eigenvalues.

Next consider the second term of (S-1). Note that,

$$\begin{aligned}
& Z \left(Z^T Z \right)^{-1} \frac{\partial \left(Z^T Z \right)}{\partial \rho} \left(Z^T Z \right)^{-1} Z^T \\
& = P_{n,\mathbf{s}} \frac{\partial Z}{\partial \rho} \left(Z^T Z \right)^{-1} Z^T + Z \left(Z^T Z \right)^{-1} \frac{\partial Z^T}{\partial \rho} P_{n,\mathbf{s}}.
\end{aligned}$$

Recall, for two square matrices A and B, AB and BA have same eigenvalues. Thus eigenvalues of $P_{n,\mathbf{s}} \frac{\partial Z}{\partial \rho} \left(Z^T Z \right)^{-1} Z^T$ is same as eigenvalues of $\frac{\partial Z}{\partial \rho} \left(Z^T Z \right)^{-1} Z^T P_{n,\mathbf{s}} = \frac{\partial Z}{\partial \rho} \left(Z^T Z \right)^{-1} Z^T$, and eigenvalues of $Z \left(Z^T Z \right)^{-1} \frac{\partial Z^T}{\partial \rho} P_{n,\mathbf{s}}$ is same as that of $P_{n,\mathbf{s}} Z \left(Z^T Z \right)^{-1} \frac{\partial Z^T}{\partial \rho} = Z \left(Z^T Z \right)^{-1} \frac{\partial Z^T}{\partial \rho}$, as $P_{n,\mathbf{s}} Z = Z$.

Finally, note that for any matrix A, $\lambda_{\max}(A + A^T)$ is bounded above by $2\lambda_{\max}(A)$, and $\lambda_{\min}(A + A^T)$ is bounded below by $2\lambda_{\min}(A)$. Hence the second term of (S-1) has finite eigenvalues. Similarly, the sum of first and third term of (S-1) has bounded eigenvalues.

Combining the above results, note that the RHS of (S-1) can written as sum of two symmetric matrices, and eigenvalues of each of them are bounded. Using the result that for symmetric matrices A and B,
$\lambda_{\min}(A) + \lambda_{\min}(B) \leq \lambda_{\min}(A+B) \leq \lambda_{\max}(A+B) \leq \lambda_{\max}(A) + \lambda_{\max}(B)$,
the result follows. □

References

Bühlmann, P. (2013). Statistical Significance in High-Dimensional Linear Models. *Bernoulli*, 19(4):1212–1242.

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-Dimensional Statistics With a View Toward Applications in Biology. *Annual Review of Statistics and Its Application*, 1(1):255–278.

Das, M. and Bhattacharya, S. (2019). Transdimensional Transformation Based Markov Chain Monte Carlo. *Brazilian Journal of Probability and Statistics*, 33(1):87–138.

Dey, K. K. and Bhattacharya, S. (2016). On Geometric Ergodicity of Additive and Multiplicative Transformation Based Markov Chain Monte Carlo in High Dimensions. *Brazilian Journal of Probability and Statistics*, 30:570–613.

Dutta, S. (2012). Multiplicative Random Walk Metropolis-Hastings on the Real Line. *Sankhya. Series B*, 74:315–342. Also available at arXiv.

Dutta, S. and Bhattacharya, S. (2014). Markov Chain Monte Carlo Based on Deterministic Transformations. *Statistical Methodology*, 16:100–116. Also available at http://arxiv.org/abs/1106.5850. Supplement available at http://arxiv.org/abs/1306.6684.

Javanmard, A. and Montanari, A. (2014). Confidence Intervals and Hypothesis Testing for High-Dimensional Regression. *Journal of Machine Learning Research*, 15:2869–2909.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). Mixtures of g Priors for Bayesian Variable Selection. *Journal of the American Statistical Association*, 103(481):410–423.

Liu, J. S. and Sabatti, S. (2000). Generalized Gibbs Sampler and Multigrid Monte Carlo for Bayesian Computation. *Biometrika*, 87:353–369.

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). p-Values for High-Dimensional Regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Roy, S. and Bhattacharya, S. (2020). Function Optimization with Posterior Gaussian Derivative Process. arXiv:2010.13591v1.

Stroeker, R. (1983). Approximations of the eigenvalues of the covariance matrix of a first-order autoregressive process. *Journal of Econometrics*, 22(3):269–279.

Wang, B. and Zhang, F. (1992). Some Inequalities for the Eigenvalues of the Product of Positive Semidefinite Hermitian Matrices. *Linear Algebra and its Applications*, 160:113–118.

Zellner, A. and Siow, A. (1980). Posterior Odds Ratios for Selected Regression Hypotheses. In Bernardo, J. M., DeGroot, M. H., Lindley, D. V., and Smith, A. F. M., editors, *Bayesian Statistics: Proceedings of the First International Meeting Held in Valencia*, pages 585–603, Valencia, Spain. University of Valencia Press.