Annals of the Institute of Statistical Mathematics (2022) 74:581–613
https://doi.org/10.1007/s10463-021-00810-6

Bayes factor asymptotics for variable selection in the Gaussian process framework

Minerva Mukhopadhyay[1] **· Sourabh Bhattacharya**[2]

Received: 31 May 2021 / Accepted: 18 August 2021 / Published online: 20 September 2021

Abstract
We investigate Bayesian variable selection in models driven by Gaussian processes, which allows us to treat linear, nonlinear and nonparametric models, in conjunction with even dependent setups, in the same vein. We consider the Bayes factor route to variable selection, and develop a general asymptotic theory for the Gaussian process framework in the "large p, large n" settings even with $p \gg n$, establishing almost sure exponential convergence of the Bayes factor under appropriately mild conditions. The fixed p setup is included as a special case. To illustrate, we apply our result to variable selection in linear regression, Gaussian process model with squared exponential covariance function accommodating the covariates, and a first-order autoregressive process with time-varying covariates. We also follow up our theoretical investigations with ample simulation experiments in the above regression contexts and variable selection in a real, riboflavin data consisting of 71 observations and 4088 covariates. For implementation of variable selection using Bayes factors, we develop a novel and effective general-purpose transdimensional, transformation-based Markov chain Monte Carlo algorithm, which has played a crucial role in simulated and real data applications.

Keywords Strong consistency · Kullback–Leibler divergence · Integrated Bayes factor · Squared exponential kernel · MCMC · Variable selection

✉ Minerva Mukhopadhyay
minervam@iitk.ac.in

Sourabh Bhattacharya
bhsourabh@gmail.com

[1] Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur 208016, India

[2] Interdisciplinary Statistical Research Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India

1 Introduction

The importance of variable selection is undeniable, since most statistical procedures involve a large number of observed variables, or covariates, only a few of which are expected to have significant influence on the experiment and future prediction. It is thus important to judiciously select those few important covariates from a relatively large pool of available covariates. This task involves multiple challenges. Even in the simple classical linear regression setup, false inclusion or exclusion of the variables may lead to false inclusion or exclusion of correlated variables. That, in turn, can influence the variance of predictions and hence the root mean square error (RMSE), and the bias of predictions. The most popular methods developed in the classical paradigm, the penalty-based methods such as the Akaike Information Criterion and the Bayesian Information Criterion, are not immune to these problems, the former having the ill reputation of preferring models consisting of relatively large number of variables. Although the latter employs a more appropriate penalty and is preferable, in practice, it can lead to underfitting. The popular LASSO method (see, for example, Tibshirani 1996) often has the effect of drastically reducing RMSE, but at the cost of increasing prediction errors. See Heinze et al. (2018) for a relatively recent review regarding several of these issues; see also Draper and Smith (2005), Weisberg (2005). Asymptotic theory of the variable selection criterion in multiple regression has been considered in Nishii (1996) and Shao (1997); see also Eubank (1999) and Giraud (2015) for various issues regarding variable selection in linear models. Since even for simple linear regression models the variable selection issues can be of significant concern, it is well imaginable how grave the issues can be in the case of more realistically complex models such as nonlinear and nonparametric regression.

Apart from some of the issues touched upon, all the classical methods of variable selection have the major drawback of selecting a single set of variables without quantifying the uncertainty associated with such selection. This calls for the Bayesian paradigm of variable selection, which is also rich in its repertoire of philosophies and methodologies. One philosophy is Bayesian model averaging, which recommends a mixture of all possible models for better prediction (see Fragoso et al. 2018 for a review). Another philosophy is to infer from the posterior distribution of the regression coefficients (see, for e.g., Ishwaran and Rao 2005). An alternative philosophy is to obtain the posterior distribution of the subsets of the covariates, and form a single posterior that encapsulates all the relevant information. Covariate selection in this case proceeds by stochastic search variable selection methods, which often involve variable-dimensional Markov chain Monte Carlo (MCMC) procedures (see O'Hara and Sillanpää 2009 for a review). Even though these methods are usually computationally demanding, most of them avoid the problems faced by the classical variable selection ideas. For details regarding various ideas on Bayesian model and variable selection along with relevant computational strategies, see, for example, Gilks and Roberts (1996), DiCiccio et al. (1997), Han and Carlin (2001), Fernández et al. (2001), Moreno and Girón (2008), Casella et al. (2009), Ando (2010), Bayarri

et al. (2012), Johnson and Rossell (2012), Hong and Preston (2012), Marin et al. (2014), Dawid and Musio (2015). Asymptotic theories on Bayesian variable selection can be found in Moreno et al. (2010), Shang and Clayton (2011), Moreno et al. (2015), Mukhopadhyay et al. (2015), although most of these theories are developed in the linear regression setup.

However, perhaps the most principled way of comparing the subsets of covariates is offered by Bayes factors, through the ratio of the posterior and prior odds associated with the competing models, which follows directly from the coherent procedure of Bayesian hypothesis testing of preferring one model compared to other. The idea is also closely related to the aforementioned principle of obtaining posterior distributions of the covariate subsets. For a general account of Bayes factors and its numerous advantages, see, for example, Kass and Raftery (1995). However, careless use of Bayes factors can lead to selecting the more parsimonious but wrong model in large samples even in very simple setups for ill-chosen priors, as the well-known Jeffreys-Lindley-Bartlett paradox demonstrates (see Jeffreys 1939; Lindley 1957; Bartlett 1957; Robert 1993; Villa and Walker 2015 for details). It is thus of utmost importance to carefully investigate the asymptotic theory of Bayes factors in different setups and construct appropriate priors that ensure consistency in the sense that the Bayes factor selects the correct set of covariates asymptotically. Note that priors that ensure consistency of posterior distributions need not guarantee consistency of Bayes factors, which is again demonstrated by the Jeffreys-Lindley-Bartlett and information paradox (see, for example, Section 2.3 of Liang et al. (2008)). Thus, the asymptotic theory of Bayes factors does not follow from the asymptotic theory of posterior distributions.

Compared to the asymptotic theory of posterior distributions, that of Bayes factors for general model selection have seen relatively slow development. Indeed, most of the theory for variable selection using Bayes factors have hitherto concentrated around nested linear regression models; see, for example, Guo and Speckman (1998), Liang et al. (2008), Moreno et al. (2010), Rousseau and Choi (2012), Wang and Sun (2014), Kundu and Dunson (2014), Choi and Rousseau (2015). However, see also Wang and Maruyama (2016) for a non-nested setup. This seems to be a very restrictive setup for the Bayesian framework, particularly in light of the current advancement in research on highly complex physical phenomena, where simplistic models are untenable. For a general account of advancements in the area of Bayes factor asymptotics, see Chib and Kuffner (2016), which also asserts the same fact.

Although variable selection has been considered in nonlinear and nonparametric frameworks such as generalized linear models, generalized additive models, additive partial linear models, generalized additive partial linear models, semiparametric additive partial linear models, additive nonparametric regression models (see, for example, Chen et al. 1999; Huang et al. 2010; Liu et al. 2011; Marra and Wood 2011; Meyer and Laud 2002; Ntzoufras et al. 2003; Reich et al. 2009; Shively et al. 1999; Wang et al. 2011; Wang and George 2007; Banerjee and Ghosal 2014), Bayes factor is not the selection criterion for the existing approaches.

It is thus crucially important to build appropriate asymptotic theory for Bayes factors with respect to variable selection in general setups. Recognizing this requirement, our endeavor in this paper is to establish consistency of Bayes factors for variable selection

in models driven by Gaussian processes. The Gaussian process framework enables us to consider linear and nonlinear, parametric, as well as nonparametric models including appropriate dependence structures, under the same umbrella, allowing the usage of a general body of mathematical apparatus to establish our asymptotic theory. Encouragingly, such a treatment allowed us to guarantee almost sure exponential convergence of the Bayes factor in favour of the true set of covariates under reasonably mild, verifiable assumptions, not only as the sample size increases indefinitely, but also as the number of available covariates increase with the sample size, possibly at faster rates, defining the so-called large p, large n paradigm, which also includes the fixed p situation as a special case. We are not aware of any asymptotic theory of Bayes factors in the "large p, large n" scenario.

We follow up the general Bayes factor convergence result with both theoretical and simulation-based illustrations of variable selection in linear regression, Gaussian process regression with squared exponential covariance function, and a first-order autoregressive model consisting of time-varying covariates.

The rest of this paper is structured as follows. We introduce the general setup for Bayes factor-based variable selection in Sect. 2. Section 3 shows almost sure convergence of the Bayes factor of any model with respect to the true model. Section 4 provides illustrations of our main result in linear regression and Gaussian process model with squared exponential covariance function. In Sect. 5, we generalize the results of Sect. 3 to the case with unknown error variance. Section 6 provides further generalization of the result, assuming arbitrary priors on compact spaces for all other parameters and hyperparameters. In Sect. 7 we treat the case of correlated errors and present the problem of time-varying covariate selection in a first-order autoregressive model as an illustration, establishing almost sure exponential convergence of the relevant Bayes factor. The important case of misspecification is dealt with in Sect. 8, where almost sure exponential convergence of Bayes factor in favor of selection of the best possible subset of covariates, is established. In Sect. 9 an overview of our simulation and real data experiments are provided; complete details are relegated to the supplement. Finally, we make concluding remarks, and provide future directions in Sect. 10.

2 General setup for Bayes factor variable selection

Let y_i and $\mathbf{x}_i$ denote the i-th response variable and the associated vector of covariates, $i = 1, \ldots, n$. We assume that the predictor $\mathbf{x}$ consists of p (> 1) components, or covariates, and that it is required to select a subset of the p components that best explains the response variable y. We allow p to grow with n at a rate $p = O(n^r)$, $r > 0$.

Let $\mathbf{s}$ denote any subset of the indices $\mathbf{S} = \{1, 2, \ldots, p\}$, and $\mathbf{x}_\mathbf{s}$ denote the co-ordinates of $\mathbf{x}$ associated with $\mathbf{s}$. To relate $\mathbf{x}_\mathbf{s}$ to y we consider the following nonparametric regression setup:

$$y = f(\mathbf{x}_\mathbf{s}) + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ is the random error and the function $f(\cdot)$ is considered unknown. We assume that $f : \mathfrak{X} \mapsto \mathrm{IR}$, where $\mathfrak{X} = \cup_{\ell=1}^{p} \mathrm{IR}^\ell$.

By assuming this framework we include the possibility that the domain of f can range from one to p-dimensions. We further assume that there exists a true set of regressors, $\mathbf{x}_0$, which influences the dependent variable y. Our problem is to identify $\mathbf{x}_0$. Note that we do not consider any specific form of the function. Irrespective of the functional form, we are only interested in identifying the set of active regressors $\mathbf{x}_0$.

2.1 The Gaussian process prior

We assign a Gaussian process prior on $f(\cdot)$ which leads, for any given subset $\mathbf{s}$ and covariate values $\left\{\mathbf{x}_{i,\mathbf{s}}; i = 1, \ldots, n\right\}$, to the joint multivariate normal distribution of $\left(f(\mathbf{x}_{1,\mathbf{s}}), \ldots, f(\mathbf{s}_{n,\mathbf{s}})\right)^T$ with mean and variance-covariance matrix as follows:

$$\begin{aligned} \boldsymbol{\mu}_{n,\mathbf{s}} &= \left(\mu(\mathbf{x}_{1,\mathbf{s}}), \ldots, \mu(\mathbf{x}_{n,\mathbf{s}})\right)^T; \\ \Sigma_{n,\mathbf{s}} &= \left(\left(Cov\left(f(\mathbf{x}_{i,\mathbf{s}}), f(\mathbf{x}_{j,\mathbf{s}})\right)\right); \; i = 1, \ldots, n; j = 1, \ldots, n. \end{aligned} \tag{2}$$

The marginal distribution of $\mathbf{y}_n = (y_1, \ldots, y_n)^T$ is then the n-variate normal,

$$\mathbf{y}_n \sim N_n\left(\boldsymbol{\mu}_{n,\mathbf{s}}, \sigma_\epsilon^2 I_n + \Sigma_{n,\mathbf{s}}\right),$$

where I_n is the identity matrix of order n. We denote this marginal model by $\mathcal{M}_\mathbf{s}$. It will be increasingly evident as we proceed, that this relatively simple consideration is the key to unlocking a sufficiently general asymptotic theory of Bayes factors for variable selection that allows handling of wide range of situations including parametric, nonparametric, independence and dependence, using the same basic concept and mathematical manoeuvre.

2.2 The true model

We assume that there exists exactly one particular subset $\mathbf{s}_0$ of $\mathbf{S}$ which is actually associated with the data generating process of y, which is termed as the *true* subset. The evaluation procedure of the proposed set of model selection basically rests on its ability to identify this true subset, irrespective of the form of the function f.

We denote the mean vector and the covariance matrix of the Gaussian process prior associated with the true model by $\boldsymbol{\mu}^t_{n,\mathbf{s}_0}$ and $\Sigma^t_{n,\mathbf{s}_0}$, respectively, and denote the corresponding marginal distribution of $\mathbf{y}_n$ as $\mathcal{M}^t_{\mathbf{s}_0}$. For notational convenience we drop the suffix n from $\boldsymbol{\mu}_{n,\mathbf{s}}$, $\boldsymbol{\mu}^t_{n,\mathbf{s}_0}$, $\Sigma_{n,\mathbf{s}}$ and $\Sigma^t_{n,\mathbf{s}_0}$.

2.3 The Bayes factor for covariate selection

It follows from the general model setup and the Gaussian process prior that the Bayes factor of any model $\mathcal{M}_\mathbf{s}$ to the true model $\mathcal{M}^t_{\mathbf{s}_0}$ associated with the data is given by

$$BF^n_{\mathbf{s},\mathbf{s}_0} = \frac{\mathcal{M}_{\mathbf{s}}(\mathbf{y}_n)}{\mathcal{M}^t_{\mathbf{s}_0}(\mathbf{y}_n)} = \frac{\left|\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\right|^{-1/2}}{\left|\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\right|^{-1/2}} \times \frac{\exp\left\{-\left(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\right)^T\left(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\right)/2\right\}}{\exp\left\{-\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)/2\right\}}, \tag{3}$$

which is the ratio of the marginal likelihoods of the observed data $\mathbf{y}_n$, under $\mathcal{M}_{\mathbf{s}}$ to $\mathcal{M}^t_{\mathbf{s}_0}$. This is the same as the ratio of the posterior odds and prior odds for $\mathbf{s}$ and $\mathbf{s}_0$, for any prior on the models. If the models for $\mathbf{s}$ and $\mathbf{s}_0$ have the same prior distribution, then (3) is the same as the posterior odds.

The aim of this paper is to establish that (3) converges to zero exponentially fast as $n \to \infty$, if $\mathbf{s} \neq \mathbf{s}_0$. We shall begin with known σ^2_ϵ and other parameters, but will subsequently generalize our theory when such quantities are unknown, and almost arbitrary, albeit sensible priors, are assigned to them. In the next section we establish the almost sure convergence of the log-Bayes factor.

3 Almost sure convergence of the log-Bayes factor

In this section we investigate Bayes factor consistency of Gaussian process regression in strong sense. We will show that for $\mathbf{s} \neq \mathbf{s}_0$, there exists an $\omega_{\mathbf{s}} \in [0, 1]$, and $\delta_{\mathbf{s}} > 0$ such that

$$\limsup_n n^{-(1+2r\omega_{\mathbf{s}})} \log BF^n_{\mathbf{s},\mathbf{s}_0} \overset{a.s.}{=} -\delta_{\mathbf{s}}.$$

The quantities $\omega_{\mathbf{s}}$, for $\mathbf{s} \subseteq \mathbf{S}$, as we shall make precise in the applications, is related to the sparsity conditions of the underlying model $\mathbf{s}_0$ and the competing model $\mathbf{s}$. One way to interpret $\omega_{\mathbf{s}}$ is to set $O(p^{\omega_{\mathbf{s}}}) = O(n^{r\omega_{\mathbf{s}}})$ as the difference in effective dimensionality of the true model $\mathbf{s}_0$ and competing model $\mathbf{s}$. Thus, when the effective dimensionality of the models indexed by $\mathbf{s}$ and $\mathbf{s}_0$ is bounded, as $n \to \infty$, then $\omega_{\mathbf{s}} = 0$. Note that depending upon the value of $\omega_{\mathbf{s}}$, we can compare models of different dimensionalities.

We first state the assumptions under which the result holds.

(A1) Let $\Delta_{n,\mathbf{s}} \overset{def}{=} (\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0})^T\left(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\right)^{-1}(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0})$. We assume that for any $\mathbf{s} \subseteq \mathbf{S}$, for some $\omega_{\mathbf{s}} \in [0, 1]$ and $\xi_{\mathbf{s}} > 0$,

$$\liminf_n n^{-(1+2r\omega_{\mathbf{s}})}\Delta_{n,\mathbf{s}} = \xi_{\mathbf{s}}.$$

Define $A_{n,\mathbf{s}} = \left(\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\right)\left(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\right)^{-1}$. We further assume the following:

(A2) Let $\lambda_1 \geq \cdots \geq \lambda_n > 0$ be the eigenvalues of $A_{n,\mathbf{s}}$, then for $\omega_{\mathbf{s}}$ defined in (A1), $\lambda_{\max}(A_{n,\mathbf{s}}) = O\left(p^{2\omega_{\mathbf{s}}}\right) = O\left(n^{2r\omega_{\mathbf{s}}}\right)$.

(*A*3) Finally we assume that for all $\mathbf{s}$, and for $\omega_{\mathbf{s}}$ defined in (A1),

$$\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\|^2 = O\big(n^{1+b}p^{2\omega_{\mathbf{s}}}\big) = O\big(n^{1+b+2r\omega_{\mathbf{s}}}\big), \text{ for some } b < 1/2.$$

We will show that, the quantity $\Delta_{n,\mathbf{s}}$ in (A1) is asymptotically equivalent to the Kullback-Leibler (KL) divergence between the marginal density of $\mathbf{y}_n$ under $\mathbf{s}$ and that under $\mathbf{s}_0$, in most of the frameworks including linear model. Thus requiring (A1) is same as requiring positive KL divergence between $\mathcal{M}_{\mathbf{s}}$ and $\mathcal{M}^t_{\mathbf{s}_0}$ after proper scaling. Assumptions (A2) and (A3) are reasonable and verifiable restrictions.

In the illustrations with linear and Gaussian process regression, we will show that $p^{\omega_{\mathbf{s}}}$ can be interpreted essentially as the cardinality of set difference of $\mathbf{s}$ and $\mathbf{s}_0$. Further, in the illustration with a first-order autoregressive model, we demonstrate that $p^{\omega_{\mathbf{s}}}$ may be interpreted essentially as $\max\big\{|\mathbf{s}|, |\mathbf{s}_0|\big\}$.

Our first result shows that limit supremum of the expected log Bayes factor of any model and the true model is negative, when scaled by $n^{1+2r\omega_{\mathbf{s}}}$.

Result 1 Assume (*A*1) holds for some $\omega_{\mathbf{s}} \in [0, 1]$. Then for some $\delta_{\mathbf{s}} > 0$ depending upon $\mathbf{s}$ ($\neq \mathbf{s}_0$),

$$\limsup_{n\to\infty} E_{\mathbf{s}_0}\Big(\frac{1}{n^{1+2r\omega_{\mathbf{s}}}} \log BF^n_{\mathbf{s},\mathbf{s}_0}\Big) = -\delta_{\mathbf{s}},$$

for the same choice of $\omega_{\mathbf{s}}$ as given in (A1).

Proof From (3) we find that the expectation of log Bayes factor is given by

$$\begin{aligned} E_{\mathbf{s}_0}\Big[\log\Big(BF^n_{\mathbf{s},\mathbf{s}_0}\Big)\Big] =& \frac{1}{2}\log\frac{\big|\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\big|}{\big|\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\big|} \\ &- \frac{1}{2}E_{\mathbf{s}_0}\Big[\big(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\big)^T\big(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\big)^{-1}\big(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\big)\Big] \\ &+ \frac{1}{2}E_{\mathbf{s}_0}\Big[\Big(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\Big)^T\Big(\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\Big)^{-1}\Big(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\Big)\Big]. \end{aligned} \tag{4}$$

To evaluate the first part in the above equation, note that

$$\frac{1}{2}\log\frac{\big|\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}\big|}{\big|\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\big|} = \frac{1}{2}\log\big|A_{n,\mathbf{s}}\big| = \frac{1}{2}\sum_{j=1}^{n}\log\lambda_j(A_{n,\mathbf{s}}).$$

For the second term of (4) we obtain

$$\begin{aligned} &E_{\mathbf{s}_0}\Big[\big(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\big)^T\big(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\big)^{-1}\big(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}\big)\Big] \\ &= tr(A_{n,\mathbf{s}}) + \Big(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\Big)^T\big(\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}\big)^{-1}\Big(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\Big) = tr(A_{n,\mathbf{s}}) + \Delta_{n,\mathbf{s}}. \end{aligned}$$

The last term of (4) is given by $E_{\mathbf{s}_0}\left[\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(\sigma^2_\epsilon I_n+\Sigma^t_{\mathbf{s}_0}\right)^{-1}\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)\right]=n$. Using the above facts and from (4) observe that

$$2E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)\right]+\Delta_{n,\mathbf{s}}=\sum_{i=1}^{n}\left(\log\lambda_i-\lambda_i+1\right).$$

Note that $g(x)=\log x-x+1$ is an increasing function on $(0,\,1]$ and decreasing function on $(1,\infty)$, having maximum at 0. Thus $\sum_i\left(\log\lambda_i-\lambda_i+1\right)\le 0$.

Thus, combining the above facts and (A1) we write

$$\limsup_n\frac{1}{n^{1+2r\omega_{\mathbf{s}}}}E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)\right]\le-\xi_{\mathbf{s}}/2.$$

Hence, there exists $\delta_{\mathbf{s}}>0$ such that $\limsup_{n\to\infty}E_{\mathbf{s}_0}\left(\frac{1}{n^{1+2r\omega_{\mathbf{s}}}}\log BF^n_{\mathbf{s},\mathbf{s}_0}\right)=-\delta_{\mathbf{s}}$. □

Next we will prove L_4 convergence of $\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)/n^{1+2r\omega_{\mathbf{s}}}$ towards its expectation, which in turn would imply L_2 convergence.

Let $B_{\mathbf{s}_0}$ be the appropriate matrix associated with the Cholesky factorization of $\sigma^2_\epsilon I_n+\Sigma^t_{\mathbf{s}_0}$, i.e., $\sigma^2_\epsilon I_n+\Sigma^t_{\mathbf{s}_0}=B_{\mathbf{s}_0}B^T_{\mathbf{s}_0}$, and $C_{n,\mathbf{s}}=B^T_{\mathbf{s}_0}\left(\sigma^2_\epsilon I_n+\Sigma_{\mathbf{s}}\right)^{-1}B_{\mathbf{s}_0}$. Then $\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}=B_{\mathbf{s}_0}z_n$, with $z_n\sim N_n\left(\mathbf{0},I_n\right)$. Then

$$\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(\sigma^2_\epsilon I_n+\Sigma^t_{\mathbf{s}_0}\right)^{-1}\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)=z_n^Tz_n,$$

and $\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(\sigma^2_\epsilon I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)=z_n^TC_{n,\mathbf{s}}z_n$. Note further that $A_{n,\mathbf{s}}$ and $C_{n,\mathbf{s}}$ have the same eigenvalues. Thus, by assumption (A2), $\lambda_{\max}(C_{n,\mathbf{s}})=O\left(n^{2r\omega_{\mathbf{s}}}\right)$.

Result 2 Assume $(A2)$ and $(A3)$ hold for some $\omega_{\mathbf{s}}\in[0,1]$. Then

$$n^{-1-2r\omega_{\mathbf{s}}}\left\{\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)-E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)\right]\right\}\xrightarrow{a.s.}0,\ \text{as}\ n\to\infty.$$

Proof For convenience, we write $\tilde{E}_n:=E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)\right]$. Now note that for $A_{n,\mathbf{s}}$, z_n, $B_{\mathbf{s}_0}$ and $C_{n,\mathbf{s}}$ as defined above

$$\begin{aligned}
&2E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right)-\tilde{E}_n\right]^4\\
=&E_{\mathbf{s}_0}\left[-z_n^TC_{n,\mathbf{s}}z_n+E_{\mathbf{s}_0}\left(z_n^TC_{n,\mathbf{s}}z_n\right)+2z_n^TB^T_{\mathbf{s}_0}\left(\sigma^2_\epsilon I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}}-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)+z_n^Tz_n-n\right]^4\\
\le&C\left[E_{\mathbf{s}_0}\left|z_n^TC_{n,\mathbf{s}}z_n-tr(C_{n,\mathbf{s}})\right|^4+E_{\mathbf{s}_0}\left|z_n^TB^T_{\mathbf{s}_0}\left(\sigma^2_\epsilon I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}}-\boldsymbol{\mu}^t_{\mathbf{s}_0}\right)\right|^4\right.\\
&\left.+E_{\mathbf{s}_0}\left|z_n^Tz_n-n\right|^4\right]
\end{aligned}\tag{5}$$

where C is a positive constant. The above result follows by repeated application of the inequality $(a+b)^q \le 2^{q-1}(a^q+b^q)$, for non-negative a, b, where $q \ge 1$.

We first obtain the asymptotic order of the first term of (5). Note that for any n vector z_n and any $n \times n$ matrix C_n

$$\begin{aligned} E\{z_n^T C_n z_n - E_{\mathbf{s}_0}(z_n^T C_n z_n)\}^4 &= E(z_n^T C_n z_n)^4 \\ &- 4E(z_n^T C_n z_n)^3 E(z_n^T C_n z_n) + 6E(z_n^T C_n z_n)^2 \{E(z_n^T C_n z_n)\}^2 \\ &- 4E(z_n^T C_n z_n)\{E(z_n^T C_n z_n)\}^3 + \{E(z_n^T C_n z_n)\}^4. \end{aligned} \tag{6}$$

To evaluate (6), we make use of the following results (see, for example, Magnus 1978; Kendall and Stuart 1947).

$$\begin{aligned} E_{\mathbf{s}_0}(z_n^T C_{n,\mathbf{s}} z_n) &= tr(C_{n,\mathbf{s}}); \\ E_{\mathbf{s}_0}(z_n^T C_{n,\mathbf{s}} z_n)^2 &= [tr(C_{n,\mathbf{s}})]^2 + 2tr\left(C_{n,\mathbf{s}}^2\right); \\ E_{\mathbf{s}_0}(z_n^T C_{n,\mathbf{s}} z_n)^3 &= [tr(C_{n,\mathbf{s}})]^3 + 6tr(C_{n,\mathbf{s}})tr\left(C_{n,\mathbf{s}}^2\right) + 8tr\left(C_{n,\mathbf{s}}^3\right); \\ E_{\mathbf{s}_0}(z_n^T C_{n,\mathbf{s}} z_n)^4 &= [tr(C_{n,\mathbf{s}})]^4 + 32tr(C_{n,\mathbf{s}})tr\left(C_{n,\mathbf{s}}^3\right) + 12\left[tr\left(C_{n,\mathbf{s}}^2\right)\right]^2 \\ &\quad + 12[tr(C_{n,\mathbf{s}})]^2 tr\left(C_{n,\mathbf{s}}^2\right) + 48tr\left(C_{n,\mathbf{s}}^4\right). \end{aligned}$$

Substituting the above expressions in (6) we obtain

$$E_{\mathbf{s}_0}\{z_n^T C_{n,\mathbf{s}} z_n - E_{\mathbf{s}_0}(z_n^T C_{n,\mathbf{s}} z_n)\}^4 = 12\left[tr\left(C_{n,\mathbf{s}}^2\right)\right]^2 + 48tr\left(C_{n,\mathbf{s}}^4\right).$$

If $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $C_{n,\mathbf{s}}$, then $\lambda_1^k, \ldots, \lambda_n^k$ are the eigenvalues of $C_{n,\mathbf{s}}^k$, for $k \in \mathbb{N}$. Therefore the above quantity reduces to

$$12\left(\sum_i \lambda_i^2\right)^2 + 48\sum_i \lambda_i^4 \le Cn\sum_i \lambda_i^4 = O\left(n^{2+8r\omega_{\mathbf{s}}}\right), \tag{7}$$

due to the fact that $\lambda_{\max}(C_{n,\mathbf{s}}) = \lambda_{\max}(A_{n,\mathbf{s}})$, (A2) and as $\left(\sum_{i=1}^n a_i\right)^2 \le n\sum_{i=1}^n a_i^2$.

Let us now obtain the asymptotic order of second term of (5). Note that, the random variable $z_n^T B_{\mathbf{s}_0}^T(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right)$ is univariate normal with mean zero and variance

$$\begin{aligned} \hat{\sigma}_n^2 =& \left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right)^T(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1}\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}_0}^t\right)(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right) \\ \le& \lambda_{\max}\left[(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1/2}\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}_0}^t\right)(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1/2}\right] \\ &\times \left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right)^T(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}})^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right) \\ \le& \sigma_\epsilon^{-2}\lambda_{\max}(A_{n,\mathbf{s}})\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\|^2 = O\left(n^{1+b+4r\omega_{\mathbf{s}}}\right), \end{aligned}$$

due to $(A2)$ and $(A3)$. Hence it follows that

$$E_{\mathbf{s}_0}\left|z_n^T B_{\mathbf{s}_0}^T\left(\sigma_\epsilon^2 I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\boldsymbol{\mu}_{\mathbf{s}}-\boldsymbol{\mu}_{\mathbf{s}_0}^t\right)\right|^4=3\hat{\sigma}_n^4=O\left(n^{2+2b+8r\omega_{\mathbf{s}}}\right). \tag{8}$$

Finally, we deal with the third term of (5). As $z_n^T z_n - n = \sum_{i=1}^n \left(z_i^2-1\right)$, where, for $i=1,\ldots,n$, $z_i^2 \overset{iid}{\sim} \chi_1^2$. By Lemma B of Serfling (1980, p. 68), it follows that

$$E_{\mathbf{s}_0}\left(z_n^T z_n-n\right)^4=O\left(n^2\right).$$

Substituting (7), (8) and the above in (5) we obtain

$$E_{\mathbf{s}_0}\left[\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)-\tilde{E}_n\right]^4=O\left(n^{2+2b+8r\omega_{\mathbf{s}}}\right). \tag{9}$$

Chebychev's inequality, in conjunction with (9) guarantees that for any $\eta>0$,

$$\sum_{n=1}^{\infty} P_{\mathbf{s}_0}\left(\left|\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)-\tilde{E}_n\right|>n^{1+2r\omega_{\mathbf{s}}}\eta\right)<\infty,$$

as $b<1/2$, proving almost sure convergence of $n^{-1-2r\omega_{\mathbf{s}}}\left\{\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)-\tilde{E}_n\right\}$ to 0, as $n\to\infty$. □

Now we state the main theorem, the proof of which follows as an application of the above result, and Result 1.

Theorem 1 (**Main theorem**) *Suppose the assumptions (A1)–(A3) hold for some* $\omega_{\mathbf{s}}\in[0,1]$, *and* $\delta_{\mathbf{s}}>0$ *depending upon* $\mathbf{s}\,(\neq\mathbf{s}_0)$, *then*

$$\limsup_n \frac{1}{n^{1+2r\omega_{\mathbf{s}}}}\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)\overset{a.s.}{=}-\delta_{\mathbf{s}}.$$

Remark 1 Recall that $p^{\omega_{\mathbf{s}}}$ is related to the effective dimensionality of the models $\mathbf{s}$ and $\mathbf{s}_0$. When p is fixed, then $p^{\omega_{\mathbf{s}}}$ is zero. Indeed, keeping p fixed and proceeding exactly in the same way as the proof of Theorem 1, and setting $\omega_{\mathbf{s}}=0$ in assumptions $(A1)$–$(A3)$, would yield the result:

$$\limsup_n n^{-1}\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)\overset{a.s.}{=}-\delta_{\mathbf{s}}.$$

Further, if p were fixed, then the number in the class of all competing models, 2^p-1, would be finite. In that case, under assumptions $(A1)$–$(A3)$ (with $\omega_{\mathbf{s}}=0$ for all $\mathbf{s}\in\mathbf{S}$), there would exist $\delta>0$, such that

$$\max_{\mathbf{s}\neq\mathbf{s}_0}\limsup_n \frac{1}{n}\log\left(BF_{\mathbf{s},\mathbf{s}_0}^n\right)\overset{a.s.}{=}-\delta.$$

Remark 2 One can establish a relatively weaker version of consistency result,

$$\limsup_n \frac{1}{n^{1-\epsilon+2r\omega_{\mathbf{s}}}} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}\right) \overset{a.s.}{=} -\delta_{\mathbf{s}}, \qquad \epsilon < 1/4,$$

under a weaker variant of assumption (A1), $(A1^\star)$: $\liminf_n n^{-1+\epsilon-2r\omega_{\mathbf{s}}}\Delta_{n,\mathbf{s}} = \xi_{\mathbf{s}}$. However, assumption (A3) should be replaced by $(A3^\star)$: $\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\|^2 = O\left(np^{2\omega_{\mathbf{s}}}\right) = O\left(n^{1+2r\omega_{\mathbf{s}}}\right)$, for all $\mathbf{s}$, which, nonetheless, remains a mild assumption. When $r\omega_{\mathbf{s}}$ is large, this version of consistency becomes more appropriate than the traditional one.

Remark 3 Theorem 1 remains valid for nested models $\mathcal{M}_{\mathbf{s}}$ and $\mathcal{M}_{\mathbf{s}_0}$ where one model has $|\mathbf{s}\Delta\mathbf{s}_0| = O(p^{\omega_{\mathbf{s}}})$ covariates more than the other, where $\omega_{\mathbf{s}} \in [0, 1]$.

4 Illustrations

This section provides illustrations of our main result in two different contexts, linear regression and Gaussian process regression with squared exponential covariance function.

4.1 Linear regression

For illustration of the Bayes factor theory we first consider the linear regression. Let $y_i = \boldsymbol{\beta}^T_{\mathbf{s}}\mathbf{x}_{i,\mathbf{s}} + \epsilon_i$, where $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma^2_\epsilon\right)$, for $i = 1, \ldots, n$. Let $\mathbf{s}_0\,(\subseteq \mathbf{S} = \{1, 2, \ldots, p\})$ be the set of indices of the true set of covariates, and $p = O(n^r)\,(r > 0)$. We assign a normal prior on $\boldsymbol{\beta}_{\mathbf{s}}$, $\boldsymbol{\beta}_{\mathbf{s}} \sim N\left(\boldsymbol{\beta}_{0,\mathbf{s}}, g_n\sigma^2_{\boldsymbol{\beta}}\left(X^T_{\mathbf{s}}X_{\mathbf{s}}\right)^{-1}\right)$, which is similar to the well known Zellner's g prior. Zellner's g-prior assigns $\boldsymbol{\beta}_{0,\mathbf{s}} = \mathbf{0}$. We instead make the prior more flexible by assuming that $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = \sum_{j=1}^{|\mathbf{s}|} |\beta_{0,j}| = O(|\mathbf{s}|)$ for all $\mathbf{s}$. We further assume that $g_n = O(p^{\omega_{\mathbf{s}}})$.

We assume that the space of covariates is compact, which, as we show, is sufficient to ensure $(A1)$–$(A3)$. Observe that Zellner's g-prior induces a Gaussian process prior on the function $f(\mathbf{x}_{i,\mathbf{s}}) = \mathbf{x}^T_{i,\mathbf{s}}\boldsymbol{\beta}_{\mathbf{s}}$ with mean function

$$\mu\left(\mathbf{x}_{i,\mathbf{s}}\right) = \boldsymbol{\beta}^T_{0,\mathbf{s}}\mathbf{x}_{i,\mathbf{s}} = \boldsymbol{\mu}_{\mathbf{s}},$$

and the covariance between $\boldsymbol{\beta}^T_{\mathbf{s}}\mathbf{x}_{i,\mathbf{s}}$ and $\boldsymbol{\beta}^T_{\mathbf{s}}\mathbf{x}_{j,\mathbf{s}}$ is given by

$$Cov\left(\boldsymbol{\beta}^T_{\mathbf{s}}\mathbf{x}_{i,\mathbf{s}}, \boldsymbol{\beta}^T_{\mathbf{s}}\mathbf{x}_{j,\mathbf{s}}\right) = \sigma^2_{\boldsymbol{\beta}}g_n\mathbf{x}^T_{i,\mathbf{s}}\left(X^T_{\mathbf{s}}X_{\mathbf{s}}\right)^{-1}\mathbf{x}_{j,\mathbf{s}}.$$

Therefore, $\Sigma_{\mathbf{s}} = \sigma^2_{\boldsymbol{\beta}}g_nX^T_{\mathbf{s}}\left(X^T_{\mathbf{s}}X_{\mathbf{s}}\right)^{-1}X_{\mathbf{s}} = \sigma^2_{\boldsymbol{\beta}}g_nP_{n,\mathbf{s}}$, where $P_{n,\mathbf{s}}$ is the projection matrix on the space of $X_{\mathbf{s}}$.

We verify assumptions (A1)–(A3) under this setup. To see that assumption (A1) holds, we first calculate the Kullback-Leibler divergence between the marginal density of $\mathbf{y}_n$ under $\mathbf{s}$ and that under $\mathbf{s}_0$, $\mathcal{KL}^n(\mathbf{s}, \mathbf{s}_0)$, which is

$$\begin{aligned}\mathcal{KL}^n(\mathbf{s},\mathbf{s}_0) &\propto tr\big(A_{n,\mathbf{s}}\big) - \log|A_{n,\mathbf{s}}| - n \\ &\quad + \left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T \left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)^{-1} \left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right) \\ &= tr\big(A_{n,\mathbf{s}}\big) - \log|A_{n,\mathbf{s}}| - n + \Delta_{n,\mathbf{s}}.\end{aligned}$$

As the eigenvalues of a projection matrix can only be zero or one, and the traces of $P_{n,\mathbf{s}_0}$ and $P_{n,\mathbf{s}}$ are $|\mathbf{s}_0|$ and $|\mathbf{s}|$, respectively, we have

$$\begin{aligned}tr\big(A_{n,\mathbf{s}}\big) &= tr\left[\left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)^{-1}\left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}_0}\right)\right] \\ &= tr\left[I_n + \sigma_\beta^2 g_n \left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)^{-1}\left(P_{n,\mathbf{s}_0} - P_{n,\mathbf{s}}\right)\right] \\ &= n + \sigma_\beta^2 g_n tr\big(D_{n,\mathbf{s}}\big),\end{aligned} \tag{10}$$

where $D_{n,\mathbf{s}} = \left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)^{-1}\left(P_{n,\mathbf{s}_0} - P_{n,\mathbf{s}}\right)$. By Result S-2 (see Section S-7 of the Supplement Material) we have

$$\left(\sigma_\epsilon^2\right)^{-1} I_n \geq \left(\sigma_\epsilon^2 I_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)^{-1} \geq \left(\sigma_\epsilon^2 + g_n\sigma_\beta^2\right)^{-1} I_n,$$

so that

$$\frac{|\mathbf{s}_0| - |\mathbf{s}|}{\sigma_\epsilon^2 + g_n\sigma_\beta^2} \leq tr\big(D_{n,\mathbf{s}}\big) \leq \frac{|\mathbf{s}_0| - |\mathbf{s}|}{\sigma_\epsilon^2}.$$

Substituting the above in (10) yields

$$n + \sigma_\beta^2 g_n \frac{|\mathbf{s}_0| - |\mathbf{s}|}{\sigma_\epsilon^2 + g_n\sigma_\beta^2} \leq tr\big(A_{n,\mathbf{s}}\big) \leq n + \sigma_\beta^2 g_n \frac{|\mathbf{s}_0| - |\mathbf{s}|}{\sigma_\epsilon^2}.$$

As $|\mathbf{s}_0| - |\mathbf{s}| \leq |\mathbf{s}\Delta\mathbf{s}_0|$, assuming that $|\mathbf{s}\Delta\mathbf{s}_0| = O(p^{\omega_{\mathbf{s}}}) = O(n^{r\omega_{\mathbf{s}}})$, in conjunction with the assumption that $g_n = O(p^{\omega_{\mathbf{s}}})$, as $n \to \infty$,

$$\frac{tr\big(A_{n,\mathbf{s}}\big)}{n^{1+2r\omega_{\mathbf{s}}}} \to \begin{cases} 1 & \text{if} \quad \omega_{\mathbf{s}} = 0; \\ 0 & \text{if} \quad \omega_{\mathbf{s}} \in (0,1]. \end{cases}$$

Further,

$$|A_{n,\mathbf{s}}| = \frac{\left|I + \sigma_\beta^2 g_n \sigma_\epsilon^{-2} P_{n,\mathbf{s}_0}\right|}{\left|I + \sigma_\beta^2 g_n \sigma_\epsilon^{-2} P_{n,\mathbf{s}}\right|} = \left(1 + \frac{\sigma_\beta^2 g_n}{\sigma_\epsilon^2}\right)^{|\mathbf{s}_0| - |\mathbf{s}|}.$$

Therefore,

$$\frac{1}{n^{1+2r\omega_{\mathbf{s}}}} \log|A_{n,\mathbf{s}}| = \frac{|\mathbf{s}_0| - |\mathbf{s}|}{n^{1+2r\omega_{\mathbf{s}}}} \log\left(1 + \sigma_\beta^2 g_n/\sigma_\epsilon^2\right) / \to 0, \text{ as } n \to \infty.$$

Combining the above facts, we get, for all $\omega_{\mathbf{s}} \in [0,1]$,

$$n^{-(1+2r\omega_{\mathbf{s}})}\left\{tr\left(A_{n,\mathbf{s}}\right) - \log|A_{n,\mathbf{s}}| - n\right\} \to 0, \text{ as } n \to \infty.$$

Thus, $\liminf_n n^{-(1+2r\omega_{\mathbf{s}})}\mathcal{KL}^n(\mathbf{s},\mathbf{s}_0) = \liminf_n n^{-(1+2r\omega_{\mathbf{s}})}\Delta_{n,\mathbf{s}}$.

The assumption (A1) is thus implied by $\liminf_n n^{-(1+2r\omega_{\mathbf{s}})}\mathcal{KL}^n(\mathbf{s},\mathbf{s}_0) > 0$, which is a natural assumption. Bounded, positive eigenvalues of $\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}}$, along with Result S-2 imply that $\Delta_{n,\mathbf{s}}$ is of the same order as $\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\|^2$, which again, is of order $O(n^{1+2r\omega_{\mathbf{s}}})$, as we show below. Viewing the requirement of $(A1)$ from this perspective, it seems natural to demand that the mean functions of the competing and the true models be distinct in the sense that $\liminf_n \|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{0,\mathbf{s}_0}\|^2/n^{1+2r\omega_{\mathbf{s}}} > 0$.

To check assumption $(A2)$ note that for positive definite Hermitian matrices A and B, $\lambda_{\max}(AB) \le \lambda_{\max}(A)\lambda_{\max}(B)$. Using this fact and as $\sigma_{\boldsymbol{\beta}}^2 g_n \sigma_\epsilon^{-2} = O(p^{\omega_{\mathbf{s}}})$, it is easily seen that

$$\lambda_{\max}(A_{n,\mathbf{s}}) \le \left(1 + \sigma_{\boldsymbol{\beta}}^2 g_n \sigma_\epsilon^{-2}\right) = O(p^{\omega_{\mathbf{s}}}).$$

Finally we check $(A3)$. Note that $\left\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{0,\mathbf{s}_0}\right\|^2 \le \left\|X_{\mathbf{s}\Delta\mathbf{s}_0}\boldsymbol{\beta}_{0,\mathbf{s}\Delta\mathbf{s}_0}\right\|^2$, as the prior mean of the j-th regression coefficient $\beta_{0,j}$ remains the same accross different models which include the j-th covariate, x_j.

Further, recall that for any $\mathbf{s}$, $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = O(|\mathbf{s}|)$. Since the covariates lie on a compact space, it follows that $\left\|X_{\mathbf{s}\Delta\mathbf{s}_0}\boldsymbol{\beta}_{0,\mathbf{s}\Delta\mathbf{s}_0}\right\|^2 = \sum_{i=1}^n \left(\mathbf{x}^T_{i,\mathbf{s}\Delta\mathbf{s}_0}\boldsymbol{\beta}_{0,\mathbf{s}\Delta\mathbf{s}_0}\right)^2 = O\left(np^{2\omega_{\mathbf{s}}}\right) = O(n^{1+2r\omega_{\mathbf{s}}})$, if $|\mathbf{s}\Delta\mathbf{s}_0| = O(p^{\omega_{\mathbf{s}}})$. Thus $(A3)$ holds.

Thus Theorem 1 holds for the linear regression setup. This result is summarized in the form of the following theorem.

Theorem 2 *Consider the linear regression model* $y_i = \boldsymbol{\beta}_{\mathbf{s}}^T\mathbf{x}_{i,\mathbf{s}} + \epsilon_i$, *where* $\epsilon_i \overset{iid}{\sim} N\left(0,\sigma_\epsilon^2\right)$, *for* $i = 1,\ldots,n$. *Let* $\boldsymbol{\beta}_{\mathbf{s}} \sim N\left(\boldsymbol{\beta}_{0,\mathbf{s}}, g_n\sigma_{\boldsymbol{\beta}}^2\left(X_{\mathbf{s}}^T X_{\mathbf{s}}\right)^{-1}\right)$, *where* $1 \le |\mathbf{s}| \le p$ *and* $p = O(n^r)$, $r > 0$. *Assume that the space of covariates is compact, and* $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = \sum_{j=1}^{|\mathbf{s}|}|\beta_{0,j}| = O(|\mathbf{s}|)$. *Further, if there exists some* $\omega_{\mathbf{s}} \in [0,1]$ *such that* $|\mathbf{s}\Delta\mathbf{s}_0| = O(p^{\omega_{\mathbf{s}}})$, *and* $\mathcal{KL}^n(\mathbf{s},\mathbf{s}_0)/(n^{1+2r\omega_{\mathbf{s}}}) > 0$, *then for* $g_n = O(p^{\omega_{\mathbf{s}}})$ *the statement of Theorem 1 holds.*

4.2 Gaussian process with squared exponential kernel

We now consider the problem of variable selection in nonparametric model of the form $y = \mathbf{x}_{\mathbf{s}}^T\boldsymbol{\beta}_{\mathbf{s}} + f(\mathbf{x}_{\mathbf{s}}) + \epsilon$, where f belongs to a Hilbert space $\mathcal{H}$. Let $f(\mathbf{x}_{\mathbf{s}})$ be modeled by a zero-mean Gaussian process with squared exponential covariance kernel of the form

$$Cov\big(f(\mathbf{x}_\mathbf{s}), f(\mathbf{x}'_\mathbf{s})\big) = \sigma_f^2 \exp\left\{-\frac{1}{2}\big(\mathbf{x}_\mathbf{s} - \mathbf{x}'_\mathbf{s}\big)^T D_\mathbf{s}\big(\mathbf{x}_\mathbf{s} - \mathbf{x}'_\mathbf{s}\big)\right\}. \tag{11}$$

Here σ_f^2 can be interpreted as the process variance, and the diagonal elements of $D_\mathbf{s}$ can be interpreted as the smoothness parameters. As in the case of linear regression, we consider the Zellner's g-prior for $\boldsymbol{\beta}_\mathbf{s}$. Thus, the mean function $\mu_\mathbf{s}$ here is of the same form as in the linear regression case. The (i, j)-th element of the covariance matrix $\Sigma_\mathbf{s}$ is given by

$$\sigma_{\boldsymbol{\beta}}^2 g_n \mathbf{x}_{i,\mathbf{s}}^T \big(X_\mathbf{s}^T X_\mathbf{s}\big)^{-1} \mathbf{x}_{j,\mathbf{s}} + \sigma_f^2 \exp\left\{-\frac{1}{2}\big(\mathbf{x}_{i,\mathbf{s}} - \mathbf{x}_{j,\mathbf{s}}\big)^T D_\mathbf{s}\big(\mathbf{x}_{i,\mathbf{s}} - \mathbf{x}_{j,\mathbf{s}}\big)\right\}.$$

True model. As before we indicate a particular subset of $\mathbf{S} = \{1, \ldots, p\}$ as the true set of regressors $\mathbf{s}_0$. The corresponding mean vector and variance matrices are denoted by $\boldsymbol{\mu}_{\mathbf{s}_0}^t$ and $\Sigma_{\mathbf{s}_0}^t$, respectively.

Assumption. Before verifying assumptions (A1)-(A3), we state the following assumption on the design matrix.

(A4) We assume that $\{\mathbf{x}_{j,\mathbf{s}} : j = 1, 2, \ldots\}$ and $D_\mathbf{s}$ are such that for all $i \geq 1$,

$$\sum_{j \neq i = 1}^{n} \exp\left\{-\frac{1}{2}\big(\mathbf{x}_{i,\mathbf{s}} - \mathbf{x}_{j,\mathbf{s}}\big)^T D_\mathbf{s}\big(\mathbf{x}_{i,\mathbf{s}} - \mathbf{x}_{j,\mathbf{s}}\big)/2\right\} = K_\mathbf{s} = O(1),$$

where $K_\mathbf{s}$ (> 0) may depend upon $\mathbf{s}$.

Verification of the assumptions. We verify assumptions (A1)–(A3) under this setup and assuming (A4) holds.

First observe that (A3) is satisfied in the same way as in the linear regression case. Before verifying (A1)-(A2), note that by Gerschgorin's circle theorem, every eigenvalue λ of any $n \times n$ matrix A with (i, j)-th element a_{ij} satisfies $|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|$, for at least one $i \in \{1, \ldots, n\}$ (see, for example, Lange 2010). It then follows by (A4) that the maximum eigenvalue of the covariance matrix associated with $f(\cdot)$ is bounded above by $K_\mathbf{s}$. Also, the covariance matrix associated with the linear part $\mathbf{x}_\mathbf{s}^T \boldsymbol{\beta}_\mathbf{s}$, being a projection matrix, has maximum eigenvalue 1. Hence, by Result S-2, we conclude that the maximum eigenvalue of $\Sigma_\mathbf{s}$ is bounded above by finite $\tilde{K}_\mathbf{s} > 0$.

To verify $(A1)$, note that $\big(\sigma_\epsilon^2 I_n + \Sigma_\mathbf{s}\big)^{-1} \succ \big(\sigma_\epsilon^2 + \tilde{K}_\mathbf{s}\big)^{-1} I_n$ by the first part of Result S-2. Hence,

$$n^{-1-2r\omega_\mathbf{s}} \Delta_{n,\mathbf{s}} > \big(\sigma_\epsilon^2 + \tilde{K}_\mathbf{s}\big)^{-1} n^{-1-2r\omega_\mathbf{s}} \|\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\|^2. \tag{12}$$

Now, if we wish to enforce distinguishability of only the mean functions of the competing models in the sense that

$$\liminf_n n^{-1-2r\omega_\mathbf{s}} \|\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\|^2 > 0,$$

then it is clear from (12) that $(A1)$ holds.

Next we check (A2). As before, it can be shown that the maximum eigenvalue of $\Sigma_{\mathbf{s}_0}$ is bounded above by $\tilde{K}_{\mathbf{s}_0}$ for some constant $\tilde{K}_{\mathbf{s}_0} > 0$. Then

$$\lambda_1(A_{n,\mathbf{s}}) = \lambda_{\max}\left[\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}}\right)^{-1}\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}_0}^t\right)\right]$$
$$\leq \frac{\lambda_{\max}\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}_0}^t\right)}{\lambda_{\min}\left(\sigma_\epsilon^2 I_n + \Sigma_{\mathbf{s}}\right)} \leq \frac{\sigma_\epsilon^2 + \tilde{K}_{\mathbf{s}_0}}{\sigma_\epsilon^2} = O(1),$$

showing that $(A2)$ holds.

Therefore, we have established the following theorem:

Theorem 3 *Consider the regression model* $y_i = \boldsymbol{\beta}_{\mathbf{s}}^T \mathbf{x}_{i,\mathbf{s}} + f(\mathbf{x}_{i,\mathbf{s}}) + \epsilon_i$, *where* $\epsilon_i \overset{iid}{\sim} N\left(0, \sigma_\epsilon^2\right)$ *for* $i = 1, \ldots, n$, *and* $\mathbf{s} \subseteq \mathbf{S} = \{1, \ldots, p\}$ *with* $p = O(n^r)$, $r > 0$. *Let* $\boldsymbol{\beta}_{\mathbf{s}} \sim N\left(\boldsymbol{\beta}_{0,\mathbf{s}}, g_n \sigma_\beta^2 \left(X_{\mathbf{s}}^T X_{\mathbf{s}}\right)^{-1}\right)$ *with* $g_n = O(p^{\omega_{\mathbf{s}}})$, *and* $f(\cdot)$ *be a zero-mean Gaussian process with a squared exponential covariance kernel of the form (11). Assume that the space of covariates is compact, and* $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = O(|\mathbf{s}|)$. *If there exists some* $\omega_{\mathbf{s}} \in [0, 1]$ *such that* $|\mathbf{s} \Delta \mathbf{s}_0| = O(p^{\omega_{\mathbf{s}}})$, *and* $\liminf_n n^{-1-2r\omega_{\mathbf{s}}} \|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}\|^2 > 0$, *and further if (A4) holds, then the statement of Theorem 1 holds.*

Additionally, consider the following remarks.

Remark 4 The condition in (12) also implies that $\liminf_n n^{-1-2r\omega_{\mathbf{s}}} \mathcal{KL}^n(\mathbf{s}, \mathbf{s}_0) = \liminf_n n^{-1-2r\omega_{\mathbf{s}}} \left\{tr(A_{n,\mathbf{s}}) - \log|A_{n,\mathbf{s}}| - n + \Delta_{n,\mathbf{s}}\right\} > 0$, since $tr(A_{n,\mathbf{s}}) - \log|A_{n,\mathbf{s}}| - n \geq 0$. Recall that in the linear regression setup as well we had replaced the KL-divergence $\liminf_n n^{-1-2r\omega_{\mathbf{s}}} \mathcal{KL}^n(\mathbf{s}, \mathbf{s}_0) > 0$ with the above mean divergence condition (12) to verify $(A1)$, since the eigenvalues of $\Sigma_{\mathbf{s}}$ in that setup are also bounded.

Remark 5 The linear regression term in the mean function can be replaced by any function $\boldsymbol{\mu}_{\mathbf{s}}$ subject to the condition $\|\boldsymbol{\mu}_{\mathbf{s}}\|_{L_1} = O(np^{2\omega_{\mathbf{s}}}) = O(n^{1+2r\omega_{\mathbf{s}}})$, where $0 \leq \omega_{\mathbf{s}} \leq 1$. It is easy to verify assumptions (A1) and (A3) under the aforementioned restriction on $\boldsymbol{\mu}_{\mathbf{s}}$.

5 The case with unknown error variance

So far we have assumed that the error variance σ_ϵ^2 is known. In reality, this may also be unknown and we need to assign a prior on the same. For our purpose, for any $\mathbf{x}_i, \mathbf{x}_i \in \mathfrak{X}$, we now set $Cov(f(\mathbf{x}_i), f(\mathbf{x}_j)) = \sigma_\epsilon^2 c(\mathbf{x}_i, \mathbf{x}_j)$, where $c(\mathbf{x}, \mathbf{y})$ is some appropriate correlation function, $i, j = 1, \ldots, n$. Thus, we set the process variance of $f(\cdot)$ to be the same as the error variance. Although this might seem somewhat restrictive from the inference perspective, for Bayes factor-based variable selection this is quite appropriate, as we establish almost sure exponential convergence of the resultant Bayes factor associated with this prior, in favour of the true set of covariates.

With the aforementioned modification, we assign the conjugate inverse-gamma prior on σ_ϵ^2 with parameters α, β as follows:

$$\pi\left(\sigma_\epsilon^2\right) = \frac{\beta^\alpha}{\Gamma(\alpha)}\sigma_\epsilon^{-2(\alpha+1)}\exp\left(-\frac{\beta}{\sigma_\epsilon^2}\right), \quad \alpha > 2,\ \beta > 0. \tag{13}$$

Under the same prior setup on f, the marginal of $\mathbf{y}_n = (y_1, \ldots, y_n)^T$ given σ_ϵ^2 is the n-variate normal, given by

$$\mathbf{y}_n \sim N_n\left(\boldsymbol{\mu}_\mathbf{s}, \sigma_\epsilon^2\left(I_n + \Sigma_\mathbf{s}\right)\right),$$

where $\Sigma_\mathbf{s}$ is as given in (2). After marginalizing σ_ϵ^2 the marginal of $\mathbf{y}_n$ is

$$m_\mathbf{s}\left(\mathbf{y}_n\right) \propto \left|I + \Sigma_\mathbf{s}\right|^{-1/2}\left\{\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right)^T\left(I_n + \Sigma_\mathbf{s}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right) + 2\beta\right\}^{-(\alpha+n/2)+1},$$

which is proportional to the density of multivariate t distribution with location parameter $\boldsymbol{\mu}_\mathbf{s}$, covariance matrix $\beta\left(I_n + \Sigma_\mathbf{s}\right)/(\alpha-1)$, and degrees of freedom $2(\alpha-1)$. Thus, $E(\mathbf{y}_n) = \boldsymbol{\mu}_\mathbf{s}$, and $Var(\mathbf{y}_n) = \beta\left(I_n + \Sigma_\mathbf{s}\right)/(\alpha-2)$, under $\mathcal{M}_\mathbf{s}$.

Here the Bayes factor of any model $\mathbf{s}$ to the true model $\mathbf{s}_0$ is

$$BF^n_{\mathbf{s},\mathbf{s}_0} = \frac{\left|I_n + \Sigma^t_{\mathbf{s}_0}\right|^{1/2}}{\left|I_n + \Sigma_\mathbf{s}\right|^{1/2}} \times \left[\frac{\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right)^T\left(I_n + \Sigma_\mathbf{s}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right) + 2\beta}{\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n + \Sigma^t_{\mathbf{s}_0}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right) + 2\beta}\right]^{-(\alpha+n/2)+1}.$$

As before, define $z_n \sim N(\mathbf{0}, I_n)$ such that

$$\begin{aligned}
z_n^T z_n &= \left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n + \Sigma^t_{\mathbf{s}_0}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right),\\
\Delta_{n,\mathbf{s}} &= (\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}^t_{\mathbf{s}_0})^T\left(I_n + \Sigma_\mathbf{s}\right)^{-1}(\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}^t_{\mathbf{s}_0}),\\
A_{n,\mathbf{s}} &= \left(I_n + \Sigma^t_{\mathbf{s}_0}\right)\left(I_n + \Sigma_\mathbf{s}\right)^{-1},\\
C_{n,\mathbf{s}} &= \left(I_n + \Sigma^t_{\mathbf{s}_0}\right)^{1/2}\left(I_n + \Sigma_\mathbf{s}\right)^{-1}\left(I_n + \Sigma^t_{\mathbf{s}_0}\right)^{1/2}, \text{ and therefore,}\\
&\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right)^T\left(I_n + \Sigma_\mathbf{s}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\right) = z_n^T C_{n,\mathbf{s}} z_n + \Delta_{n,\mathbf{s}}\\
&- 2\left(\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n + \Sigma_\mathbf{s}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right).
\end{aligned}$$

It then follows that

$$
\begin{aligned}
&\frac{1}{n\log n}\log BF^{n}_{\mathbf{s},\mathbf{s}_0}\\
&= \frac{\log|C_{n,\mathbf{s}}|}{2n\log n} - \frac{1}{\log n}\left(\frac{1-\alpha}{n}-\frac{1}{2}\right)\Bigg[\log\left(\frac{z_n^T z_n + 2\beta}{n^{1+2r\omega_{\mathbf{s}}}}\right)\\
&- \log\left\{\frac{\Delta_{n,\mathbf{s}} + 2\beta + z_n^T C_{n,\mathbf{s}} z_n}{n^{1+2r\omega_{\mathbf{s}}}} - 2\frac{\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)}{n^{1+2r\omega_{\mathbf{s}}}}\right\}\Bigg] \qquad (14)\\
&\le \frac{1}{2\log n}\log\left\{\frac{tr(C_{n,\mathbf{s}})}{n}\right\} - \frac{1}{\log n}\left(\frac{1-\alpha}{n}-\frac{1}{2}\right)\Bigg[\log\left(\frac{z_n^T z_n + 2\beta}{n}\right)\\
&\qquad -2r\omega_{\mathbf{s}}\log n - \log\left\{\frac{\Delta_{n,\mathbf{s}} + tr(C_{n,\mathbf{s}}) + 2\beta}{n^{1+2r\omega_{\mathbf{s}}}} + \frac{z_n^T C_{n,\mathbf{s}} z_n - tr(C_{n,\mathbf{s}})}{n^{1+2r\omega_{\mathbf{s}}}}\right.\\
&\qquad \left.+\frac{2}{n^{1+2r\omega_{\mathbf{s}}}}\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)\right\}\Bigg],
\end{aligned}
$$

where the last inequality is due to the log-sum inequality. We modify assumptions (A1)–(A3) by replacing σ^2_ϵ by 1, and term them $(A1')$–$(A3')$.

Next observe the following facts:

(i) $E\left[\left(z_n^T C_{n,\mathbf{s}} z_n - tr(C_{n,\mathbf{s}})\right)\right]^4 = O(n^{2+8r\omega_{\mathbf{s}}})$ implying that $\left[z_n^T C_{n,\mathbf{s}} z_n - tr(C_{n,\mathbf{s}})\right]/n^{1+2r\omega_{\mathbf{s}}} \xrightarrow{a.s.} 0$. One can prove this in exactly similar way as done in Result 2, using assumptions $(A1')$–$(A3')$.

(ii) Similarly, it can be shown that $E\left[\left(z_n^T z_n - n\right)\right]^4 = O(n^2)$ implying $z_n^T z_n/n \xrightarrow{a.s.} 1$.

(iii) From the above fact, it follows that $\log\left(z_n' z_n/n\right) \xrightarrow{a.s.} 0$ by continuous mapping theorem.

(iv) Applying $(A3')$, it can be shown that

$$
E\left[\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)\right]^4 = O(n^{2+8r\omega_{\mathbf{s}}+2b}),
$$

which in turn implies

$$
\left(\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)^T\left(I_n+\Sigma_{\mathbf{s}}\right)^{-1}\left(\mathbf{y}_n - \boldsymbol{\mu}^t_{\mathbf{s}_0}\right)/n^{1+2r\omega_{\mathbf{s}}} \xrightarrow{a.s.} 0.
$$

(v) Finally, $tr(C_{n,\mathbf{s}}) = tr(A_{n,\mathbf{s}}) \le n\lambda_{\max}(A_{n,\mathbf{s}})$, and $\lambda_{\max}(A_{n,\mathbf{s}}) = O(p^{2\omega_{\mathbf{s}}})$.

Using the above facts, it is easy to see that the right hand side of (14) has lim sup $-2r\omega_{\mathbf{s}} = -\delta_{\mathbf{s}}$, which is strictly negative when $\omega_{\mathbf{s}} \in (0, 1]$.

When $\omega_{\mathbf{s}} = 0$, similar steps as above would lead to the result

$$\limsup_n n^{-1} \log BF^n_{\mathbf{s},\mathbf{s}_0} \overset{a.s.}{=} -\delta_{\mathbf{s}}.$$

Consequently, the following result holds:

Theorem 4 *Consider the setup of Theorem 1 except that the error variance σ^2_ϵ is now unknown. Let an inverse gamma prior with parameters α and β be applied to σ^2_ϵ. Assume that (A1′)–(A3′) hold for some $\omega_{\mathbf{s}} \in (0, 1]$, and some positive constant $\delta_{\mathbf{s}}$ depending upon $\mathbf{s} (\neq \mathbf{s}_0)$. Then*

$$\limsup_n \frac{1}{n \log n} \log \left(BF^n_{\mathbf{s},\mathbf{s}_0} \right) \overset{a.s.}{=} -\delta_{\mathbf{s}}.$$

For $\omega_{\mathbf{s}} = 0$, the following holds:

$$\limsup_n \frac{1}{n} \log \left(BF^n_{\mathbf{s},\mathbf{s}_0} \right) \overset{a.s.}{=} -\delta_{\mathbf{s}}.$$

Moreover, if the number of models is finite then there exists $\delta > 0$ such that

$$\max_{\mathbf{s} \neq \mathbf{s}_0} \limsup_n \frac{1}{n} \log \left(BF^n_{\mathbf{s},\mathbf{s}_0} \right) \overset{a.s.}{=} -\delta.$$

6 Convergence of integrated Bayes factor

Let us suppose, as is usual, that the Bayes factor $BF^n_{\mathbf{s},\mathbf{s}_0}$ depends on a set of parameters and hyperparameters, denoted by $\boldsymbol{\theta}$. We denote the Bayes factor by $BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})$ instead of $BF^n_{\mathbf{s},\mathbf{s}_0}$ to indicate it's dependence on $\boldsymbol{\theta}$. If $\pi(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$, supported on $\boldsymbol{\Theta}$, then the integrated Bayes factor is given by

$$IBF^n_{\mathbf{s},\mathbf{s}_0} = \int_{\boldsymbol{\Theta}} BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

The following convergence result provides conditions under which the integrated Bayes factor converges to zero almost surely.

Theorem 5 *Consider the set up of Theorem 1 (or, the set up of Theorem 4), and assume that (A1)–(A3) (or, (A1′)–(A3′)) hold for some $\omega_{\mathbf{s}} \in [0, 1]$ and $\delta_{\mathbf{s}} > 0$, and for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and that $\boldsymbol{\Theta}$ is compact. Let $g(n) = n^{1+2r\omega_{\mathbf{s}}}$ (under the setup of Theorem 1); and $g(n) = n \log n$, or n if $\omega_{\mathbf{s}} \in (0, 1]$, or $\omega_{\mathbf{s}} = 0$, respectively, (under the setup of Theorem 4). Also assume the following:*

(i) $\log \left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}) \right) / g(n)$ *is stochastically equicontinuous,*

(ii) $E \left[\log \left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}) \right) / g(n) \right]$ *is equicontinuous with respect to $\boldsymbol{\theta}$ as $n \to \infty$, and*

(iii) *The* lim sup *and* lim inf *of* $E\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right]$ *are upper and lower semicontinuous in* $\boldsymbol{\theta}$*, respectively.*

Then, there exists $\delta_{\mathbf{s}} > 0$ *such that*

$$\limsup_n \frac{1}{g(n)} \log\left(IBF^n_{\mathbf{s},\mathbf{s}_0}\right) \overset{a.s.}{=} -\delta_{\mathbf{s}}. \tag{15}$$

Proof As assumptions $(A1)$–$(A3)$ (or $(A1')$–$(A3')$) hold by hypothesis, Theorem 1 (or Theorem 4) holds. While proving the theorem we have shown

$$\frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right) - E\left[\frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right] \xrightarrow{a.s.} 0 \quad \text{pointwise in } \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

By conditions (i) and (ii) of Theorem 5, the difference of the above two functions is stochastically equicontinuous. Further, as $\boldsymbol{\Theta}$ is compact, by the stochastic Ascoli lemma (see, e.g., Newey 1991),

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right) - E\left[\frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right] \right| \xrightarrow{a.s.} 0, \text{ as } n \to \infty.$$

In other words, given any data sequence, for any $\epsilon > 0$, there exists $n_0(\epsilon)$ such that for $n \geq n_0(\epsilon)$,

$$\left| \frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right) - E\left[\frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right] \right| < \epsilon/2, \tag{16}$$

for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Let us now define $\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}$ and $\delta_{\mathbf{s}}(\boldsymbol{\theta})$ such that
$-\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})} = \limsup_n E_{\mathbf{s}_0}\left[\log\left(\overline{BF^n_{\mathbf{s},\mathbf{s}_0}}(\boldsymbol{\theta})\right)/g(n)\right]$ and
$-\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})} = \liminf_n \underline{E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})/g(n)\right)\right]}$, where $\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}, \underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})} > 0$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.
By assumption (iii), $\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}$ is upper semicontinuous in $\boldsymbol{\theta}$ and $\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}$ is lower semicontinuous in $\boldsymbol{\theta}$.

Now, by compactness of $\boldsymbol{\Theta}$, we have $\boldsymbol{\Theta} \subset \cup_{i=1}^m \tilde{\boldsymbol{\Theta}}_i$, for some finite $m > 0$, where $\tilde{\boldsymbol{\Theta}}_i$ are such that $\sup_{\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\in\tilde{\boldsymbol{\Theta}}_i} \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\| < \delta$. Here $\delta\ (> 0)$ is such that

$$\left| E_{\mathbf{s}_0}\left\{ \frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}_1)\right)\right\} - E_{\mathbf{s}_0}\left\{ \frac{1}{g(n)} \log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}_2)\right)\right\} \right| < \frac{\epsilon}{6}, \tag{17}$$

for large n, due to equicontinuity. Now, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\boldsymbol{\theta}$ must lie in $\tilde{\boldsymbol{\Theta}}_i$ for some $i = 1, 2, \ldots, m$. Let $\boldsymbol{\theta}_i \in \tilde{\boldsymbol{\Theta}}_i$, for $i = 1, \ldots, m$. Then, let us write

$$
\begin{aligned}
&E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right\}+\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\\
&=\left[E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right\}-E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}_i)\right)\right\}\right] \qquad (18)\\
&\quad+\left[E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta}_i)\right)\right\}+\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta}_i)}\right]-\left(\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta}_i)}-\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\right).
\end{aligned}
$$

The first term on the right hand side of of (18) is less than $\epsilon/6$ due to (17), since both $\boldsymbol{\theta}, \boldsymbol{\theta}_i \in \tilde{\boldsymbol{\Theta}}_i$. The second term on the right hand side of of (18) is less than $\epsilon/6$ for large enough n by definition of lim sup. Since m is finite, the requisite $n_1(\epsilon)$ that n needs to exceed, remains finite for all values of $\boldsymbol{\theta}$. The third term is less than $\epsilon/6$ by definition of upper semicontinuity, given that $\boldsymbol{\theta}, \boldsymbol{\theta}_i \in \tilde{\boldsymbol{\Theta}}_i$. In other words, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, there exists $n_1(\epsilon)$, such that $n \geq n_1(\epsilon)$,

$$
E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right\}+\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}<\frac{\epsilon}{2}.
$$

Similarly, using the definition of equicontinuity, lim inf and lower semicontinuity, it follows that there exists $n_2(\epsilon) \geq 1$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ such that for $n \geq n_2(\epsilon)$,

$$
E_{\mathbf{s}_0}\left\{\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)\right\}+\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}>-\frac{\epsilon}{2}.
$$

From (16) and the above facts, we see that for $n \geq n_3(\epsilon) = \max\{n_1(\epsilon), n_2(\epsilon)\}$, and all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$
\begin{aligned}
&-\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}-\epsilon<\frac{1}{g(n)}\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)<-\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}+\epsilon,\\
\implies &\exp\left\{-g(n)(\epsilon+\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})})\right\}<BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})<\exp\left\{g(n)(\epsilon-\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})})\right\}.
\end{aligned}
$$

Integrating the above with respect to $\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, and taking $g(n)^{-1}\log$ we obtain,

$$
-\epsilon+\frac{1}{g(n)}\log \underline{I}_n<\frac{1}{g(n)}\log\left(IBF^n_{\mathbf{s},\mathbf{s}_0}\right)<\epsilon+\frac{1}{g(n)}\log \bar{I}_n,
$$

where $\bar{I}_n=\int_{\boldsymbol{\Theta}}\exp\left(-g(n)\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\right)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$, $\underline{I}_n=\int_{\boldsymbol{\Theta}}\exp\left(-g(n)\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\right)\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. Since both $\bar{I}_n$ and $\underline{I}_n$ are less than one, the statement of Theorem 5 holds. □

Remark 6 Note that a sufficient condition for stochastic equicontinuity of $\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)$ is almost sure Lipschitz continuity of the same, with a bounded Lipschitz constant, as $n \to \infty$. Similarly, a sufficient condition of equicontinuity of $E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right]$ is Lipschitz continuity. Again, Lipschitz continuity is ensured by boundedness of the partial derivatives. Hence, if the partial derivatives of $\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)$ and its expectation with respect to the components of $\boldsymbol{\theta}$ exist

and are almost surely bounded for large n, then Lipschitz continuity would follow. This would also imply the semicontinuity assumptions on $E_{\mathbf{s}_0}\left\{\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right\}$. In our applications, we shall often make use of this sufficient condition.

Remark 7 Note that Theorem 5 is applicable to Gaussian process regression setup where the error variance σ^2_ϵ, the process variance σ^2_f, or the diagonal elements of $D_{\mathbf{s}}$ are unknown. The relevant priors, however, need to have compact supports. Although for σ^2_ϵ and σ^2_f compactly supported prior is not necessary for proving convergence of Bayes factor (as we have shown consistency under an inverse-gamma prior setup with $\sigma^2_\epsilon = \sigma^2_f$), but very general priors, albeit with compact supports, can be envisaged for these unknown quantities, without any loss of generality of convergence result for the corresponding integrated Bayes factor. In real problems, some other parameters may be assigned compactly supported priors, while the inverse-gamma prior may be allotted to the variance parameters.

For illustration of the method for verifying the conditions of Theorem 5, in Sect. 7.1, we consider the case of variable selection in an autoregressive regression model with unknown autoregressive parameter.

7 Bayes factor asymptotics for correlated errors

So far we assumed $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$. However, correlated errors play significant roles in time series models. Indeed, except some simple cases, *i.i.d.* errors will not be appropriate for such models. For instance, the problem of time-varying covariate selection in the AR(1) model $y_t = \rho_0 y_{t-1} + \sum_{i=0}^{|\mathbf{s}|} \beta_i x_{it} + \epsilon_t$, $t = 1, 2, \ldots$, where $\epsilon_t \overset{iid}{\sim} N(0, \sigma^2_\epsilon)$ and ρ_0 is known, admits the same treatment as in linear regression considered in Sect. 4.1 by treating $z_t = y_t - \rho_0 y_{t-1}$ as the response. However if ρ_0 is unknown, such simple method is untenable.

In general, we must allow correlated errors, that is, for $\boldsymbol{\epsilon}_n = (\epsilon_1, \ldots, \epsilon_n)^T \sim N_n\left(\mathbf{0}, \sigma^2_\epsilon \tilde{\Sigma}_n\right)$, the zero-mean normal distribution with covariance matrix $\sigma^2_\epsilon \tilde{\Sigma}_n$. Let the correlation matrix under the true model be $\tilde{\Sigma}^t_n$. With these, we then replace the previous notions $\sigma^2_\epsilon I_n + \Sigma_{\mathbf{s}}$ and $\sigma^2_\epsilon I_n + \Sigma^t_{\mathbf{s}_0}$ by $\sigma^2_\epsilon \tilde{\Sigma}_n + \Sigma_{\mathbf{s}}$ and $\sigma^2_\epsilon \tilde{\Sigma}^t_n + \Sigma^t_{\mathbf{s}_0}$, respectively, and prove similar results with the assumptions on $A_{n,\mathbf{s}}$ and $\Delta_{n,\mathbf{s}}$, where $A_{n,\mathbf{s}} = \left(\sigma^2_\epsilon \tilde{\Sigma}^t_n + \Sigma^t_{\mathbf{s}_0}\right)\left(\sigma^2_\epsilon \tilde{\Sigma}_n + \Sigma_{\mathbf{s}}\right)^{-1}$, and $\Delta_{n,\mathbf{s}} = (\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0})^T \left(\sigma^2_\epsilon \tilde{\Sigma}_n + \Sigma_{\mathbf{s}}\right)^{-1} (\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0})$.

7.1 Illustration 3: autoregressive model

Consider the time-varying covariate selection problem in the following $AR(1)$ model

$$y_t = \rho y_{t-1} + \boldsymbol{\beta}'_{\mathbf{s}} \mathbf{x}_{t,\mathbf{s}} + \epsilon_t, \quad \text{and} \quad \epsilon_t \overset{iid}{\sim} N\left(0, \sigma^2_\epsilon\right), \quad \text{for} \quad t = 1, \ldots, n. \tag{19}$$

where $y_0 \equiv 0$ and $|\rho| < 1$.

The above model admits the following representation

$$y_t = \boldsymbol{\beta}_{\mathbf{s}}' z_{t,\mathbf{s}} + \tilde{\epsilon}_t, \quad \text{where } z_{t,\mathbf{s}} = \sum_{k=1}^{t} \rho^{t-k} \mathbf{x}_{k,\mathbf{s}} \quad \text{and } \tilde{\epsilon}_t = \sum_{k=1}^{t} \rho^{t-k} \epsilon_k.$$

Thus, $\tilde{\epsilon}_t$ is an asymptotically stationary zero mean Gaussian process with covariance

$$Cov\left(\tilde{\epsilon}_{t+h}, \tilde{\epsilon}_t\right) \sim \frac{\sigma_\epsilon^2 \rho^h}{1-\rho^2}, \text{ where } h \geq 0. \tag{20}$$

Let the true model be of the same form as above but with ρ and $\mathbf{s}$ replaced by ρ_0 and $\mathbf{s}_0$, respectively, where $|\rho_0| < 1$. As in the linear regression case we allow $p = O(n^r)$ covariates, with $r > 0$, and $\mathbf{s}_0 \subseteq \mathbf{S} = \{1, \ldots, p\}$.

Let $\boldsymbol{\beta}_{\mathbf{s}} \sim N\left(\boldsymbol{\beta}_{0,\mathbf{s}}, g_n \sigma_\beta^2 \left(Z_{\mathbf{s}}' Z_{\mathbf{s}}\right)^{-1}\right)$, where $Z_{\mathbf{s}}$ is the design matrix associated with $z_{t,\mathbf{s}}$; $t = 1, \ldots, n$, and $g_n = O(1)$. This is again Zellner's g prior, but modified to suit the $AR(1)$ setup.

As before, $\boldsymbol{\beta}_{0,\mathbf{s}}$ is so chosen that $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = \sum_{=1}^{|\mathbf{s}|} |\beta_{0,j}| = O(|\mathbf{s}|)$. We also assume compactness of the covariate space and that the set of covariates $\left\{x_j : j \in \mathbf{S}\right\}$ is nonzero. Let $\pi(\rho)$ be any prior for ρ supported on $[-1+\gamma, 1-\gamma]$ for some small enough $\gamma > 0$. The reason for choosing this support will become clear as we proceed.

The conditional expectation of $\boldsymbol{\beta}_{\mathbf{s}}' z_{i,\mathbf{s}}$, and the covariance between $\boldsymbol{\beta}_{\mathbf{s}}^T z_{i,\mathbf{s}}$ and $\boldsymbol{\beta}_{\mathbf{s}}^T z_{j,\mathbf{s}}$ given ρ, for $i, j = 1, \ldots, n$ are

$$\mu\left(z_{i,\mathbf{s}}\right) = \boldsymbol{\beta}_{0,\mathbf{s}}' z_{i,\mathbf{s}}, \quad \text{and} \quad Cov\left(\boldsymbol{\beta}_{\mathbf{s}}' z_{i,\mathbf{s}}, \boldsymbol{\beta}_{\mathbf{s}}' z_{j,\mathbf{s}}\right) = \sigma_\beta^2 g_n z_{i,\mathbf{s}}' \left(Z_{\mathbf{s}}' Z_{\mathbf{s}}\right)^{-1} z_{j,\mathbf{s}}.$$

Let Σ_ϵ be the AR(1) correlation matrix $((\rho^h))$, $\sigma_\epsilon^2 \tilde{\Sigma}_n$ be the covariance matrix of $\tilde{\epsilon}$ as given in (20), i.e., $\tilde{\Sigma}_n = (1-\rho^2)^{-1} \Sigma_\epsilon$, $H_{n,\mathbf{s}} := \left(\sigma_\epsilon^2 \tilde{\Sigma}_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}}\right)$ and $H_{n,\mathbf{s}_0} := \sigma_\epsilon^2 \tilde{\Sigma}_n + \sigma_\beta^2 g_n P_{n,\mathbf{s}_0}$, where $P_{n,\mathbf{s}}$ is the projection matrix onto the column space of $Z_{\mathbf{s}}$. Then $A_{n,\mathbf{s}} = H_{n,\mathbf{s}_0} H_{n,\mathbf{s}}^{-1}$.

We first verify (A1)–(A3) in this setup. For verification of (A3), note that

$$\left\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}_{\mathbf{s}_0}^t\right\|^2 = \left\|Z_{\mathbf{s}} \boldsymbol{\beta}_{0,\mathbf{s}} - Z_{\mathbf{s}_0} \boldsymbol{\beta}_{0,\mathbf{s}_0}\right\|^2 \leq 2\left(\left\|Z_{\mathbf{s}} \boldsymbol{\beta}_{0,\mathbf{s}}\right\|^2 + \left\|Z_{\mathbf{s}_0} \boldsymbol{\beta}_{0,\mathbf{s}_0}\right\|^2\right).$$

Now, by our assumptions, for any $\mathbf{s}$, $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = O(|\mathbf{s}|)$. We further assume that $\max\{|\mathbf{s}|, |\mathbf{s}_0|\} = O(p^{2\omega_{\mathbf{s}}})$, for $0 \leq \omega_{\mathbf{s}} \leq 1$. Also, since the covariates lie on a compact space and $|\rho|$ is less than one, it follows that $\left\|Z_{\mathbf{s}} \boldsymbol{\beta}_{0,\mathbf{s}}\right\|^2 = \sum_{t=1}^{n} \left(z_{t,\mathbf{s}}^T \boldsymbol{\beta}_{0,\mathbf{s}}\right)^2 = O(np^{2\omega_{\mathbf{s}}})$ $= O(n^{1+2r\omega_{\mathbf{s}}})$.

Similarly, since $|\rho_0| < 1$, $\left\|Z_{\mathbf{s}_0} \boldsymbol{\beta}_{0,\mathbf{s}}\right\|^2 = O(n^{1+2r\omega_{\mathbf{s}}})$. Thus $(A3)$ holds.

Next we verify $(A2)$. Note that, by Lemma S-1 (in Section S-7 of the supplement) the eigenvalues of $\Sigma_\epsilon/(1-\rho^2)$ have strictly positive lower and upper bounds, independent of n if $\rho \in [-1+\gamma, 1-\gamma]$. Further, the eigenvalues of $P_{n,\mathbf{s}}$ are either 0 or 1. Thus by Result S-2, $\lambda_{\max}(A_{n,\mathbf{s}}) = O(p^{\omega_{\mathbf{s}}})$.

Assuming, as before, that $\liminf_n n^{-1-2r\omega_s}\|\boldsymbol{\mu}_{\mathbf{s}} - \boldsymbol{\mu}^t_{\mathbf{s}_0}\|^2 > 0$, it is seen that $(A1)$ also holds. Thus, $(A1)$–$(A3)$ holds.

Next we verify conditions (i)–(iii) of Theorem 5. Note that

$$\begin{aligned}
&\frac{\partial}{\partial\rho}\Big(\frac{1}{n^{1+2r\omega_s}}\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)\Big)\\
&= -\frac{1}{2n^{1+2r\omega_s}}tr\Big[H^{-1}_{n,\mathbf{s}}\frac{\partial}{\partial\rho}\big(H_{n,\mathbf{s}}\big)\Big] + \frac{1}{n^{1+2r\omega_s}}\left(\frac{\partial\boldsymbol{\mu}_{\mathbf{s}}}{\partial\rho}\right)^T H^{-1}_{n,\mathbf{s}}(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}) \qquad (21)\\
&\quad + \frac{1}{2n^{1+2r\omega_s}}(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}})^T H^{-1}_{n,\mathbf{s}}\frac{\partial}{\partial\rho}\big(H_{n,\mathbf{s}}\big)H^{-1}_{n,\mathbf{s}}(\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}}).
\end{aligned}$$

Consider the first term of (21).

By Lemma S-1 $H_{n,\mathbf{s}}$ has positive and bounded eigenvalues. Define $D_{n,\mathbf{s}} = \frac{\partial}{\partial\rho}H_{n,\mathbf{s}}$, and note that

$$D_{n,\mathbf{s}} = \frac{\sigma^2_\epsilon}{1-\rho^2}A_n(\rho) + \frac{2\sigma^2_\epsilon\rho}{(1-\rho^2)^2}\Sigma_\epsilon + \sigma^2_\beta g_n\frac{\partial}{\partial\rho}P_{n,\mathbf{s}}, \qquad (22)$$

where $A_n(\rho) = ((a_{i,j}))$ is defined by $a_{i,j} = |i-j|\rho^{|i-j|-1}$. We will show that $D_{n,\mathbf{s}}$ has finite eigenvalues. From Lemma S-1 and Lemma S-2 (in Section S-7 of the supplement), and the fact that $\rho \in [-1+\gamma, 1-\gamma]$, it is evident that the 2nd and 3rd matrices in the RHS of (22) have bounded eigenvalues if g_n is bounded. As both the matrices are symmetric, it follows from Result S-2 that the sum of these two matrices have finite eigenvalues. From Gerschgorin's circle theorem, $\lambda_{\max}\big(A_n(\rho)\big) \le \max_j R_{[j]}$, and $\lambda_{\min}\big(A_n(\rho)\big) \ge -\max_j R_{[j]}$ where $R_{[j]}$ is the sum of the absolute values of the non-diagonal entries in the $[j]$-th row of $A_n(\rho)$ and $[j]$ is the highest integer less than or equal to j. Little algebra shows that $\max_j R_{[j]} = R_{[n/2]} = 2\big\{1-|\rho|^{[n/2]} - [n/2]|\rho|^{[n/2]}(1-|\rho|)\big\}(1-|\rho|)^{-2}$. As n is large $(1-|\rho|)^{-2} < R_{[n/2]} < 2(1-|\rho|)^{-2}$, which implies that the eigenvalues of the 1st matrix of RHS of (22) are bounded. Thus, $D_{n,\mathbf{s}}$ has bounded eigenvalues by Result S-2.

Let $\alpha_0 > 0$ be such that $\lambda_{\min}(D_{n,\mathbf{s}}) > -\alpha_0$. Then $D_{n,\mathbf{s}} + \alpha_0 I$ is a symmetric positive definite matrix. Hence, the absolute value on first term of (21) is

$$\begin{aligned}
\left|\frac{1}{2n^{1+2r\omega_s}}tr\Big(H^{-1}_{n,\mathbf{s}}D_{n,\mathbf{s}}\Big)\right| &= \left|\frac{1}{2n^{1+2r\omega_s}}tr\Big[H^{-1}_{n,\mathbf{s}}(D_{n,\mathbf{s}} + \alpha_0 I) - \alpha_0 H^{-1}_{n,\mathbf{s}}\Big]\right|\\
&\le \frac{1}{2n^{1+2r\omega_s}}\Big[\Big|trH^{-1}_{n,\mathbf{s}}(D_{n,\mathbf{s}} + \alpha_0 I)\Big| + \alpha_0\Big|trH^{-1}_{n,\mathbf{s}}\Big|\Big]\\
&\le \frac{1}{2n^{1+2r\omega_s}}\lambda_1\Big(H^{-1}_{n,\mathbf{s}}\Big)tr\big(D_{n,\mathbf{s}} + \alpha_0 I\big)\\
&\qquad + \frac{\alpha_0}{2n^{1+2r\omega_s}}tr\Big(H^{-1}_{n,\mathbf{s}}\Big) = O(1).
\end{aligned}$$

The last equality holds as the eigenvalues of $H_{n,\mathbf{s}}$ are positive and bounded, that of $D_{n,\mathbf{s}}$ are bounded, and α_0 is finite.

Next consider the third term of (21). Let $H_{n,\mathbf{s}}^{-1}(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}) = \mathbf{u}_\mathbf{s}$, then this term is $\mathbf{u}_{n,\mathbf{s}}^T D_{n,\mathbf{s}} \mathbf{u}_{n,\mathbf{s}} / 2n^{1+2r\omega_\mathbf{s}}$. Using Result S-2 we argue that the third term of (21) is lower bounded by $\lambda_{\min}(D_{n,\mathbf{s}})\|\mathbf{u}_{n,\mathbf{s}}\|^2/2n^{1+2r\omega_\mathbf{s}}$ and upper bounded by $\lambda_{\max}(D_{n,\mathbf{s}})\|\mathbf{u}_{n,\mathbf{s}}\|^2/2n^{1+2r\omega_\mathbf{s}}$. Using Result S-2, it can also be shown that $\|\mathbf{u}_{n,\mathbf{s}}\|^2$ is bounded by $\lambda_{\max}^{-2}(H_{n,\mathbf{s}})\|\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\|^2$ and $\lambda_{\min}^{-2}(H_{n,\mathbf{s}})\|\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\|^2$.

Next, we write $\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}_0} = H_{n,\mathbf{s}_0}^{1/2}\tilde{\boldsymbol{z}}_n$, where $\tilde{\boldsymbol{z}}_n \sim N(\mathbf{0}, I_n)$. It then follows that

$$\begin{aligned}
\|\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\|^2 &= \|\mathbf{y}_n - \boldsymbol{\mu}_{\mathbf{s}_0}^t\|^2 + \|\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}\|^2 + 2(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s})^T(\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}) \\
&\le \tilde{\boldsymbol{z}}_n^T H_{n,\mathbf{s}_0}\tilde{\boldsymbol{z}}_n + 2\|\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}\|\sqrt{\tilde{\boldsymbol{z}}_n^T H_{n,\mathbf{s}_0}\tilde{\boldsymbol{z}}_n} + \|\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}\|^2 \\
&\le \lambda_{\max}(H_{n,\mathbf{s}_0})\|\tilde{\boldsymbol{z}}_n\|^2 + 2\|\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}\|\lambda_{\max}^{1/2}(H_{n,\mathbf{s}_0})\|\tilde{\boldsymbol{z}}_n\| + \|\boldsymbol{\mu}_{\mathbf{s}_0}^t - \boldsymbol{\mu}_\mathbf{s}\|^2.
\end{aligned}$$

Combining the facts that $\lambda_{\max}(H_{n,\mathbf{s}_0})$ is bounded, $\tilde{\boldsymbol{z}}_n^T\tilde{\boldsymbol{z}}_n/n \to 1$ almost surely as $n \to \infty$, $\|\boldsymbol{\mu}_\mathbf{s} - \boldsymbol{\mu}_{\mathbf{s}_0}\|^2/n^{1+2r\omega_\mathbf{s}} = O(1)$, and since $|\rho_0|, |\rho| < 1-\gamma$, almost surely, it follows that $\|\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\|^2/n^{1+2r\omega_\mathbf{s}} = O(1)$, almost surely. In other words, the third term of (21) is $O(1)$ almost surely, as $n \to \infty$.

For the second term of (21), note that

$$\left|\left(\frac{d\boldsymbol{\mu}_\mathbf{s}}{d\rho}\right)^T H_{n,\mathbf{s}}^{-1}(\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s})\right| \le \sqrt{\left(\frac{d\boldsymbol{\mu}_\mathbf{s}}{d\rho}\right)^T\left(\frac{d\boldsymbol{\mu}_\mathbf{s}}{d\rho}\right)} \times \|\mathbf{u}_\mathbf{s}\|.$$

Note that $\frac{\partial}{\partial\rho}\boldsymbol{\mu}_\mathbf{s} = \frac{\partial}{\partial\rho}\left(\boldsymbol{\beta}_{0,\mathbf{s}}^T z_{1,\mathbf{s}}, \ldots, \boldsymbol{\beta}_{0,\mathbf{s}}^T z_{n,\mathbf{s}}\right)^T$. Thus,

$$\begin{aligned}
\left(\frac{\partial}{\partial\rho}\boldsymbol{\mu}_\mathbf{s}\right)^T\left(\frac{\partial}{\partial\rho}\boldsymbol{\mu}_\mathbf{s}\right) &= \sum_{t=1}^n\left(\frac{\partial}{\partial\rho}\boldsymbol{\beta}_{0,\mathbf{s}}^T z_{t,\mathbf{s}}\right)^2 = \sum_{t=1}^n\left(\frac{\partial}{\partial\rho}\boldsymbol{\beta}_{0,\mathbf{s}}^T\sum_{k=1}^t\rho^{t-k}\mathbf{x}_{k,\mathbf{s}}\right)^2 \\
&= \sum_{t=1}^n\left\{\sum_{k=1}^{t-1}(t-k)\rho^{t-k-1}\boldsymbol{\beta}_{0,\mathbf{s}}^T\mathbf{x}_{k,\mathbf{s}}\right\}^2 \\
&\le M_n\sum_{t=1}^n\left\{\sum_{k=1}^{t-1}(t-k)\rho^{t-k-1}\right\}^2 \\
&\le M_n\sum_{t=1}^n\left\{\frac{1-\rho^{t-1}(t-t\rho+\rho)}{(1-\rho)^2}\right\}^2,
\end{aligned}$$

for an appropriate M_n of order $O(p^{2\omega_\mathbf{s}})$ as $\mathbf{x}_{k,\mathbf{s}}$ is uniformly bounded for all k, and $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|^2 = O(\|\mathbf{s}\|^2) = O(p^{2\omega_\mathbf{s}})$. As $|\rho| < 1-\gamma$, the last expression is $O(n)$. Moreover, as $\lambda_{\max}\left[H_{n,\mathbf{s}}^{-2}\right]$ is bounded and $\|\mathbf{y}_n - \boldsymbol{\mu}_\mathbf{s}\|^2/n^{1+2r\omega_\mathbf{s}}$ is $O(1)$ almost surely, it follows that the second term of (21) is $O(1)$ almost surely.

In other words, all the three terms of (21) are $O(1)$ almost surely, as $n \to \infty$. That is, almost surely, as $n \to \infty$,

$$\frac{d}{d\rho}\left(\frac{1}{n^{1+2r\omega_\mathbf{s}}}\log BF_{\mathbf{s},\mathbf{s}_0}^n(\rho)\right) = O(1). \tag{23}$$

Therefore, for any given data sequence in the relevant non-null set, the function $\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)/n^{1+2r\omega_\mathbf{s}}$ is Lipschitz continuous in ρ. Importantly, (23) shows that there exists $n_0 \geq 1$, such that for $n \geq n_0$, the Lipschitz constant for $\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)/n^{1+2r\omega_\mathbf{s}}$ remains the same. In the same way, it can be shown that $E_{\mathbf{s}_0}\left[\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)/n^{1+2r\omega_\mathbf{s}}\right]$ is also Lipschitz in ρ, with bounded Lipschitz constant, as $n \to \infty$.

Further, assuming that the lim sup and lim inf of $E_{\mathbf{s}_0}\left[\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)/n^{1+2r\omega_\mathbf{s}}\right]$ are upper and lower semicontinuous, respectively, and appealing to Theorem 5, (15) holds.

We summarize this in the form of the following theorem.

Theorem 6 *Consider the model selection problem in the AR(1) model (19) with $p = O(n^r)$, with $r > 0$. Suppose a prior π, supported on $[-1+\gamma, 1-\gamma]$, is assigned on ρ, and ρ_0 is the true value of ρ, with $|\rho_0| < 1-\gamma$, for some $\gamma > 0$. Let $\mathbf{s}_0$, $\mathbf{s}\,(\subseteq \mathbf{S})$ be the set of indices of the true set of covariates, and a competing model. Assume that $\max\left\{|\mathbf{s}_0|, |\mathbf{s}|\right\} = O(p^{\omega_\mathbf{s}})$, for some $0 \leq \omega_\mathbf{s} \leq 1$. Let $\boldsymbol{\beta}_\mathbf{s} \sim N\left(\boldsymbol{\beta}_{0,\mathbf{s}}, g_n\sigma^2_\beta\left(Z'_\mathbf{s}Z_\mathbf{s}\right)^{-1}\right)$, where $\|\boldsymbol{\beta}_{0,\mathbf{s}}\|_{L_1} = O(p^{\omega_\mathbf{s}})$, and $g_n = O(1)$. If the space of covariates is compact, and the set of covariates $\left\{x_j : j \in \mathbf{S}\right\}$ is nonzero, then provided that the* lim sup *and* lim inf *of $E_{\mathbf{s}_0}\left[\log BF^n_{\mathbf{s},\mathbf{s}_0}(\rho)/n^{1+2r\omega_\mathbf{s}}\right]$ are upper and lower semicontinuous, respectively, (15) holds.*

Note that for simplicity we have assumed σ^2_ϵ to be known in the proof of Theorem 6. However, as the following corollary shows, this is not necessary.

Corollary 1 *Due to Theorem 4, the result of Theorem 6 continues to hold with $n^{1+2r\omega_\mathbf{s}}$ replaced with $n\log n$ if we set $\sigma^2_\beta = \sigma^2_\epsilon$ and assign the conjugate inverse-gamma prior (13) to σ^2_ϵ.*

Remark 8 Before proceeding further, it is important to understand the role of $\omega_\mathbf{s}$ in the results obtained so far. It is evident that $\omega_\mathbf{s}$ is related to the effective dimensions of $\mathcal{M}_\mathbf{s}$ and $\mathcal{M}^t_{\mathbf{s}_0}$. When the mean function of the Gaussian process, $\boldsymbol{\mu}_\mathbf{s}$, is linear (or, a smooth function of the linear combination of covariates in $|\mathbf{s}|$), and the regression coefficient of the j-th covariate has same prior mean across different models $\mathcal{M}_\mathbf{s}$ involving it, then $|\mathbf{s}\Delta\mathbf{s}_0| = O(p^{\omega_\mathbf{s}})$. This is observed in the linear and Gaussian process regression with squared exponential kernel. However, if this simplification is not available, and $\boldsymbol{\mu}_\mathbf{s}$ is any function satisfying $\|\boldsymbol{\mu}_\mathbf{s}\|^2 = O\left(n|\mathbf{s}|^2\right)$, then $\max\{|\mathbf{s}|, |\mathbf{s}_0|\} = O(p^{\omega_\mathbf{s}})$, which is observed in the AR(1) illustration. Finally, if the dimensions of the competing models do not grow with n, then $\omega_\mathbf{s} = 0$. Although the role of $\omega_\mathbf{s}$ varies with the problem's setup, existence of an $\omega_\mathbf{s}$ for which (A1)–(A3) hold, is certain. Consequently, strong Bayes factor consistency is achieved at the rate $n^{1+2r\omega_\mathbf{s}}$.

8 Variable selection using Bayes factors under misspecified situations

So far we have investigated consistency of the Bayes factor for variable selection when the true model $\mathcal{M}_{\mathbf{s}_0}$ is present in the space of models being compared. However, for a large number of covariates such an assumption need not always be realistic. Indeed, in practice, for a large number available covariates, it is usually not feasible to compare all possible models. As the true subset $\mathbf{s}_0$ is unknown, it is not unlikely to exclude it from the set of models being considered for comparison. In such cases of omissions, it makes sense to select the best subset $\mathbf{s}$ from the available class of subsets using Bayes factors. Result 3, which may be viewed as an adaptation of Theorem 5 for comparing models that are not necessarily correct, establishes the usefulness of Bayes factors even in the face of such misspecifications.

First consider a simple case. Let $\mathbf{s}_1, \mathbf{s}_2 \subseteq \mathbf{S}$ be two competing models of similar order, in the sense that either $\omega_{\mathbf{s}_1}, \omega_{\mathbf{s}_2} > 0$, or $\omega_{\mathbf{s}_1} = \omega_{\mathbf{s}_2} = 0$. The following result holds in this setup.

Result 3 Consider the setup of Sect. 6 with unknown error variance σ_ϵ^2. Let there exist $\omega_{\mathbf{s}_1}, \omega_{\mathbf{s}_2}$, such that $(A1')$–$(A3')$ hold for the models $\mathcal{M}_{\mathbf{s}_1}$ and $\mathcal{M}_{\mathbf{s}_2}$, for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\boldsymbol{\Theta}$ is compact. Assume that, $g(n) = n$ if $\omega_{\mathbf{s}_1} = \omega_{\mathbf{s}_2} = 0$, and $g(n) = n \log n$ if $\omega_{\mathbf{s}_1}, \omega_{\mathbf{s}_2} \in (0, 1]$. Also assume the following:

(i) $\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)$ is stochastically equicontinuous,
(ii) $E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right]$ is equicontinuous with respect to $\boldsymbol{\theta}$ as $n \to \infty$, and
(iii) The limit of $E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right]$ exists and is continuous in $\boldsymbol{\theta}$.

If $\mathbf{s}_1$ and $\mathbf{s}_2$ are not equal to $\mathbf{s}_0$, then there exist $\delta_{\mathbf{s}_1}, \delta_{\mathbf{s}_2} > 0$ associated with models $\mathcal{M}_{\mathbf{s}_1}$ and $\mathcal{M}_{\mathbf{s}_2}$ such that

$$\lim_n \frac{1}{g(n)} \log\left(IBF^n_{\mathbf{s}_1,\mathbf{s}_2}\right) \overset{a.s.}{=} -(\delta_{\mathbf{s}_1} - \delta_{\mathbf{s}_2}).$$

Proof Using similar arguments as in the proof of Theorem 5, under the assumptions (i)–(iii), one can show that for any $\mathbf{s}$,

$$\lim_n \frac{1}{g(n)} \log\left(IBF^n_{\mathbf{s},\mathbf{s}_0}\right) \overset{a.s.}{=} -\delta_{\mathbf{s}}(\tilde{\boldsymbol{\theta}}_{\mathbf{s}}), \tag{24}$$

where, due to (iii), $\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})} = \delta_{\mathbf{s}}(\boldsymbol{\theta}) = \delta_{\mathbf{s}}(\boldsymbol{\theta}) = \lim_n E_{\mathbf{s}_0}\left[\log\left(BF^n_{\mathbf{s},\mathbf{s}_0}(\boldsymbol{\theta})\right)/g(n)\right]$, is continuous for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $\tilde{\boldsymbol{\theta}}_{\mathbf{s}} \in \overline{\boldsymbol{\Theta}}$ such that by the mean value theorem for integrals,

$$\bar{I}_n = \int_{\boldsymbol{\Theta}} \exp\left(-g(n)\overline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\right)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \exp\left(-g(n)\delta_{\mathbf{s}}(\tilde{\boldsymbol{\theta}}_{\mathbf{s}})\right)$$
$$= \int_{\boldsymbol{\Theta}} \exp\left(-g(n)\underline{\delta_{\mathbf{s}}(\boldsymbol{\theta})}\right)\pi(\boldsymbol{\theta})d\boldsymbol{\theta} = \underline{I}_n.$$

Noting that

$$\frac{1}{g(n)}\log\left(IBF^{n}_{\mathbf{s}_1,\mathbf{s}_2}\right)=\frac{1}{g(n)}\log\left(IBF^{n}_{\mathbf{s}_1,\mathbf{s}_0}\right)-\frac{1}{g(n)}\log\left(IBF^{n}_{\mathbf{s}_2,\mathbf{s}_0}\right), \tag{25}$$

the proof is completed by taking limits of both sides of (25), applying (24) on the two terms on the right hand side, and denoting $\delta_{\mathbf{s}}(\tilde{\boldsymbol{\theta}}_{\mathbf{s}})$ by $\delta_{\mathbf{s}}$ for all $\mathbf{s}$. □

Remark 9 From Result 3 it follows that $\mathcal{M}_{\mathbf{s}_1}$ is the better model than $\mathcal{M}_{\mathbf{s}_2}$ if $\delta_{\mathbf{s}_1}<\delta_{\mathbf{s}_2}$ and $\mathcal{M}_{\mathbf{s}_2}$ is to be preferred over $\mathcal{M}_{\mathbf{s}_1}$ if $\delta_{\mathbf{s}_1}>\delta_{\mathbf{s}_2}$. The Bayes factor converges exponentially fast to infinity and zero, respectively, in these cases. Hence, asymptotically with respect to the Bayes factor, the best subset $\mathbf{s}$ is the one that minimizes $\delta_{\mathbf{s}}$.

Remark 10 Let $\omega_{\mathbf{s}_1}=0$ and $\omega_{\mathbf{s}_2}>0$. In this case, it is evident that the model $\mathcal{M}_{\mathbf{s}_1}$ is closer to the true model $\mathcal{M}_{\mathbf{s}_0}$ than $\mathcal{M}_{\mathbf{s}_2}$, in the sense that, either $|\mathbf{s}_0\Delta\mathbf{s}_1|/|\mathbf{s}_0\Delta\mathbf{s}_2|\to 0$, or $\max\{|\mathbf{s}_1|,|\mathbf{s}_0|\}/\max\{|\mathbf{s}_2|,|\mathbf{s}_0|\}\to 0$ (see Remark 8), i.e., $\mathcal{M}_{\mathbf{s}_2}$ has significantly large number of different covariates than $\mathcal{M}_{\mathbf{s}_0}$, compared to $\mathcal{M}_{\mathbf{s}_1}$. Taking $g(n)=n\log n$, and following the steps of Result 3, one can show that

$$\lim_{n}\frac{1}{g(n)}\log\left(IBF^{n}_{\mathbf{s}_2,\mathbf{s}_1}\right)\overset{a.s.}{=}-\delta_{\mathbf{s}_2}.$$

Thus, the Bayes factor favors $\mathcal{M}_{\mathbf{s}_1}$ over $\mathcal{M}_{\mathbf{s}_2}$, and converges to 0 at an exponentially fast rate.

9 An overview of our simulation and real data experiments

We consider two sets of simulation experiments. In the first set, we provide direct validation of our theoretical results by fixing a true set of covariates and comparing it with specifically chosen incorrect sets of covariates using Bayes factor as the sample size is increased. We demonstrate the validity of our results in the linear regression, Gaussian process regression, as well as in the AR(1) regression context.

In the second simulation scenario, our goal is to identify, using Bayes factors, the true set of data-generating covariates among the set of 2^p-1 available subsets of covariates, given any value of p and n. To this end, we devise a novel and efficient variable-dimensional MCMC algorithm for general-purpose variable selection using Bayes factors, in the framework of Transdimensional Transformation-based Markov Chain Monte Carlo (TTMCMC) introduced by Das and Bhattacharya (2019).

Not only do we demonstrate the effectiveness of our strategy with simulation studies involving linear, Gaussian process and AR(1) regressions, but also very successfully apply our procedure to the variable selection problem in a real riboflavin data consisting of $p=4088$ covariates and $n=71$ data points, using both linear and Gaussian process regression.

9.1 A briefing on our simulation studies for direct theory validation

In this section σ^2 is assumed to be unknown, and is assigned an inverse-gamma(1, 1) prior. The covariates are generated from scaled $t_{(3)}$ distribution, with an AR(1) structured scale matrix Σ_0, where ρ varies from 0.1–0.25. The total number of covariates p is fixed at 100, where n varies from 150 to 600. Three choices of $|\mathbf{s}_0|$ are taken, viz. $|\mathbf{s}_0| = 10, 40, 70$.

As per our result, we expect the Bayes factor of the true model against any other model to converge to zero as $n \to \infty$. We pre-select two competing models which are closest to the true model, in appropriate sense. First, a supermodel having k additional covariates, is considered. Second, we choose a model which has the same cardinality as the true model, and exactly k variables are different from the true model. For illustration 1 (linear model) and 3 (AR(1) model) we choose $k = 1$, and for illustration 2 (GP with squared exponential kernel), we choose $k = 5$. We fix the true σ^2 at 1.

We also consider the case for misspecified models in linear regression and GP regression framework. In both the cases we consider two supermodels of the true model, $\mathcal{M}_{\mathbf{s}_1}$ and $\mathcal{M}_{\mathbf{s}_2}$, having k_1 and k_2 extra covariates, and $\mathbf{s}_1 \subset \mathbf{s}_2$. Clearly, $\mathcal{M}_{\mathbf{s}_1}$ is closer to the true model than $\mathcal{M}_{\mathbf{s}_2}$. The simulation set up is kept the same as before. For linear regression we choose $k_1 = 1$, $k_2 = 5$, and for GP regression we choose $k_1 = 5$, $k_2 = 15$.

Finally, for each pair (p, n) and each example, the data-generation procedure is repeated 100 times to reduce randomness, and the mean Bayes factor is reported. Very encouraging results are obtained with our strategies in each of the regression scenarios considered. For misspecified models, it is clearly observed that Bayes factor chooses the better model, i.e., $\mathcal{M}_{\mathbf{s}_1}$, at a growing rate with n. The complete details are provided in Section S-1 of the supplement.

9.2 Simulation experiments with Bayes factor oriented TTMCMC

Although a plethora of methods are available for Bayesian variable selection (see, for example, O'Hara and Sillanpää 2009 for a review), including variable-dimensional solutions in the linear and generalized linear regression contexts (see, for example, Sillanpää and Arjas 1998; Lunn et al. 2006; Sillanpää et al. 2004; Chevalier et al. 2020), implementation of variable selection in the nonparametric Gaussian process regression setup, to the best of our knowledge, is nonexistent in the literature. Therefore, it is imperative to develop new methodologies for practical variable selection implementation in this framework.

Note that when the available number of covariates is even reasonably large, evaluation of the marginal density of the data, even if available in closed form, is infeasible to compute for all possible covariate subsets. Thus, direct comparison of all possible covariate subsets with respect to the marginal density is generally infeasible, and hence suitable MCMC approaches are necessary.

The traditional MCMC approaches are not valid in the model selection scenario. Indeed, different competing models may consist of sets of parameters with varying cardinalities, which would render the fixed-dimensional MCMC methods invalid. In the variable selection setup, at least the regression coefficients of the competing models associated with different subsets of covariates, are variable-dimensional.

Thus, variable-dimensional MCMC methods are necessary to handle the Bayesian model selection paradigm. Although reversible jump MCMC (RJMCMC) (Green 1995) is a valid model-jumping MCMC method, its effectiveness with respect to practical implementation is often very doubtful, with poor mixing properties being the integral part. Thus, considerably more innovative and effective variable-dimensional MCMC procedures are necessary to meet the challenges of complex variable-dimensional problems, such as model selection, among many others.

As such, we shall offer a generic and effective variable-dimensional, Bayes factor oriented solution to any variable selection problem. We employ the novel TTMCMC methodology of Das and Bhattacharya (2019) for general variable-dimensional problems, which is a generalization of the fixed-dimensional Transformation-based Markov Chain Monte Carlo (TMCMC) of Dutta and Bhattacharya (2014). The most important feature of TMCMC is facilitation of updating all the variables in question simultaneously using appropriate deterministic transformations of even a singleton random variable. This general strategy leads to remarkable improvement of acceptance rates and mixing properties, even in high dimensions. These key features are inherited by TTMCMC in the transdimensional context.

Here we devise a novel TTMCMC algorithm for generic variable selection problems using mixtures of additive and multiplicative transformations of singleton variables, further supplemented with another deterministic transformation step to enhance mixing. The algorithm is available as Algorithm S-2.1 in Section S-2 of the supplement. An important aspect of the algorithm is to propose a new covariate in the "birth move" by Bayes Information Criterion (BIC), given a set of existing covariates.

The method of computation of Bayes factors using TTMCMC samples is detailed in Section S-3 of the supplement. In Section S-4 of the supplement we provide the proof of its convergence.

The proposed TTMCMC strategy leads to quite effective variable selection, while exhibiting good mixing properties. We demonstrate this with simulation experiments in linear regression, Gaussian process regression and time series regression setups (see Section S-5 of the supplement).

9.3 Overview of our real data experiment

For real data application of our Bayes factor oriented variable selection procedure, we consider a dataset on riboflavin (vitamin B_2) production rate, where the response variable is the log-transformed riboflavin production rate and the covariates are the logarithms of 4088 gene expression levels. There are only $n = 71$ data points in the data (thus, a *bona fide* real example of the "large p, small n" setup). This data, made publicly available by Bühlmann et al. (2014), has been analyzed by various research

groups using traditional classical methods in the linear regression framework. We model this data as linear regression, as well as Gaussian process regression, using our Bayes factor-based covariate selection, and obtain very interesting and insightful results as compared to the existing results (see Section S-6 of SM).

10 Summary, conclusion and future direction

This work is an effort to establish an asymptotic theory of variable selection using Bayes factor in a general Gaussian process framework that encompasses linear, nonlinear, parametric, nonparametric, independent, as well as dependent setups involving a set of covariates, the size of which is allowed to increase even at much faster rates than the sample size. The setup also includes the special case where the available number of covariates is considered fixed. That even in such a general setup it has been possible to establish almost sure exponential convergence of the Bayes factor in favour of the correct subset of covariates, seems to be quite encouraging. The illustrations in the case of linear regression, Gaussian process model with squared exponential covariance function, and a first-order autoregressive model with time-varying covariates, vindicate the wide applicability of our asymptotic theory. Besides, it has been possible to adapt our main results on Bayes factor consistency to misspecified cases, where the true set of covariates is not included in the subsets of covariates to be compared using Bayes factor. As already explained, misspecification has high likelihood in practice, and from this perspective, the result on almost sure exponential convergence even for misspecifications, seems to be a pleasant one. Recalling the predominance of linear or additive model-based Bayes factor asymptotics, and "in probability" convergence of the Bayes factor, our efforts in this work attempt to provide a significant advancement.

Furthermore, we have conducted ample simulation experiments to supplement our theoretical investigations. Indeed, not only have we provided direct validation of our theoretical results; with an eye to variable selection in practical problems, we have devised a generic Bayes factor oriented TTMCMC algorithm for such purpose, demonstrating its efficacy in detecting the true set of covariates from among a very large pool (size $2^p - 1$) of available subsets of covariates, in linear, Gaussian process and AR(1) regression setups. Our TTMCMC strategy also yielded very interesting (and perhaps quite important) variable selection results in the case of a real riboflavin dataset consisting of 4088 covariates and 71 samples, exemplifying an authentic "large p, small n" real-life scenario.

It is easy to discern that our results and the methods of our proofs can be generalized without substantial modifications to situations where parts of the models are also necessary to select from among a set of possibilities, besides the best set of covariates. For example, in our linear regression example, choice might be necessary between linear and some specified nonlinear regression functions which also encapsulate the covariates in appropriate forms. In our Gaussian process example with squared exponential covariance function, the form of the covariance function may itself be questionable, and needs to be chosen from a set of plausible covariance forms, associated with various stationary and non-stationary Gaussian processes. In

the first-order autoregressive model example, the order of the autoregression may itself need to be selected. Our primary calculations confirm that our Bayes factor asymptotics admit extension to simultaneous selection of these model parts and the covariates, with additional mild assumptions. These findings, with details, will be communicated elsewhere.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10463-021-00810-6.

Acknowledgements We are sincerely grateful to the Associate Editor and the referees whose comments have led to significant improvement of our article.

References

Ando, T. (2010). *Bayesian model selection and statistical modeling*. Boca Raton, FL: CRC Press.

Banerjee, S., Ghosal, S. (2014). Bayesian variable selection in generalized additive partial linear models. *Stat, 3,* 363–378.

Bartlett, M. (1957). A comment on D. V. Lindley's statistical paradox. *Biometrika, 44*, 533–534.

Bayarri, M. J., Berger, J. O., Forte, A., García-Donato, G. (2012). Criteria for Bayesian model choice with application to variable selection. *The Annals of Statistics, 40,* 1550–1577.

Bühlmann, P., Kalisch, M., Meier, L. (2014). High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application, 1*(1), 255–278.

Casella, G., Giron, F. J., Martinez, M. L., Moreno, E. (2009). Consistency of Bayesian procedures for variable selection. *The Annals of Statistics, 37,* 1207–1228.

Chen, M.-H., Ibrahim, J. G., Yiannoutsos, C. (1999). Prior elicitation, variable selection and bayesian computation for logistic regression models. *Journal of the Royal Statistical Society. Series B, 61,* 223–242.

Chevalier, A., Fearnhead, P., Sutton, M. (2020). Reversible jump PDMP samplers for variable selection. arXiv:2010.11771v1.

Chib, S., Kuffner, T. A. (2016). Bayes factor consistency. arXiv:1607.00292.

Choi, T., Rousseau, J. (2015). A note on Bayes factor consistency in partial linear models. *Journal of Statistical Planning and Inference, 166,* 158–170.

Das, M., Bhattacharya, S. (2019). Transdimensional transformation based Markov chain Monte Carlo. *Brazilian Journal of Probability and Statistics, 33*(1), 87–138.

Dawid, A. P., Musio, M. (2015). Bayesian model selection based on proper scoring rules. *Bayesian Analysis, 10,* 479–499.

DiCiccio, T. J., Kass, R. E., Raftery, A. E., Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association, 92,* 903–915.

Draper, N. R., Smith, H. (2005). *Applied regression analysis* (3rd ed.). New York: Wiley.

Dutta, S., Bhattacharya, S. (2014). Markov chain Monte Carlo based on deterministic transformations. *Statistical Methodology, 16*, 100–116. http://arxiv.org/abs/1106.5850. Supplement available at http://arxiv.org/abs/1306.6684.

Eubank, R. (1999). *Nonparametric regression and spline smoothing* (2nd ed.). New York, NY: Marcel Dekker.

Fernández, C., E, L., Steel, M. F. J. (2001). Benchpark priors for Bayesian model averaging. *Journal of Econometrics, 100*, 381–427.

Fragoso, T. M., Bertoli, W., Louzada, F. (2018). Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review, 86*(1), 1–28.

Gilks, W. R., Roberts, G. O. (1996). Strategies for improving MCMC. In W. Gilks, S. Richardson, D. Spiegelhalter (Eds.), *Markov chain Monte Carlo in practice, interdisciplinary statistics* (pp. 89–114). London: Chapman and Hall.

Giraud, C. (2015). *Introduction to high-dimensional statistics*. Boca Raton, FL: CRC Press.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*, 711–732.
Guo, R., Speckman, P. (1998). Bayes factor consistency in linear models. In *2009 International Workshop on Objective Bayes Methodology*. Philadelphia
Han, C., Carlin, B. P. (2001). Markov chain Monte Carlo methods for computing Bayes factors. *Journal of the American Statistical Association, 96,* 1122–1132.
Heinze, G., Wallisch, C., Dunkler, D. (2018). Variable selection—review and recommendations for the practicing statistician. *Biometrical Journal, 60,* 431–449.
Hong, H., Preston, B. (2012). Bayesian averaging, prediction and nonnested model selection. *Journal of Econometrics, 167,* 358–369.
Huang, J., Horowitz, J. L., Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics, 38,* 2282–2313.
Ishwaran, H., Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics, 33*(2), 730–773.
Jeffreys, H. (1939). *Theory of probability* (1st ed.). Oxford: The Clarendon Press.
Johnson, V. E., Rossell, D. (2012). Bayesian model selection in high-dimensional settings. *Journal of the American Statistical Association, 107,* 649–660.
Kass, R. E., Raftery, R. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 773–795.
Kendall, M. G., Stuart, A. (1947). *The advanced theory of statistics* (3rd ed., Vol. I). London: Charles Griffin & Co.
Kundu, S., Dunson, D. B. (2014). Bayes variable selection in semiparametric linear models. *Journal of the American Statistical Association, 109,* 437–447.
Lange, K. (2010). *Numerical analysis for statisticians* (2nd ed.). New York: Springer.
Liang, F., Paulo, R., Molina, G., Clyde, M. A., Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. *Journal of the American Statistical Association, 103*(481), 410–423.
Lindley, D. (1957). A statistical paradox. *Biometrika, 44,* 187–192.
Liu, X., Wang, L., Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica, 21,* 1225–1248.
Lunn, D. J., Whittaker, J. C., Best, N. (2006). A Bayesian toolkit for genetic association studies. *Genetic Epidemiology, 30,* 231–247.
Magnus, J. R. (1978). The moments of products of quadratic forms in normal variables. *Statistica Neerlandica, 32,* 201–210.
Marin, J. M., Pillai, N. S., Robert, C. P., Rousseau, J. (2014). Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society. Series B, 76,* 833–859.
Marra, G., Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics and Data Analysis, 55,* 2372–2387.
Meyer, M. C., Laud, P. W. (2002). Predictive variable selection in generalized linear models. *Journal of the American Statistical Association, 97,* 859–871.
Moreno, E., Girón, F. J. (2008). Comparison of Bayesian objective procedures for variable selection in linear regression. *TEST, 17,* 472–490.
Moreno, E., Girón, F. J., Casella, G. (2010). Consistency of objective Bayes factors as the model dimension grows. *The Annals of Statistics, 38,* 1937–1952.
Moreno, E., Girón, F. J., Casella, G. (2015). Posterior model consistency in variable selection as the model dimension grows. *Statistical Science, 30,* 228–241.
Mukhopadhyay, M., Samanta, T., Chakrabarti, A. (2015). On consistency and optimality of Bayesian variable selection based on g-prior in normal linear regression models. *Annals of the Institute of Statistical Mathematics, 67,* 963–997.
Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica, 59,* 1161–1167.
Nishii, R. (1996). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics, 12*(2), 758–765.
Ntzoufras, I., Dellaportas, P., Forster, J. J. (2003). Bayesian variable and link determination for generalised linear models. *Journal of Statistical Planning and Inference, 111,* 165–180.
O'Hara, R. B., Sillanpää, M. J. (2009). A review of Bayesian variable selection methods: What, how and which. *Bayesian Analysis, 4*(1), 85–117.
Reich, B. J., Storlie, C. B., Bondell, H. D. (2009). Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics, 51,* 110–120.

Robert, C. P. (1993). A note on Jeffreys-Lindley paradox. *Statistica Sinica, 3,* 601–608.
Rousseau, J., Choi, T. (2012). Bayes factor consistency in regression problems. Unpublished report.
Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.
Shang, Z., Clayton, M. K. (2011). Consistency of Bayesian linear model selection with a growing number of parameters. *Journal of Statistical Planning and Inference, 141,* 3463–3474.
Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica, 7*, 221–264. with discussion.
Shively, T. S., Kohn, R., Wood, S. (1999). Variable selection and function estimation in additive non-parametric regression using a data-based prior. *Journal of the American Statistical Association, 94,* 777–806.
Sillanpää, M. J., Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics, 148,* 1373–1388.
Sillanpää, M. J., Gasbarra, D., Arjas, E. (2004). Comment on "on the metropolis-hastings acceptance probability to add or drop a quantitative trait locus in Markov chain Monte Carlo-based Bayesian analyses." *Genetics, 167*, 1037.
Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B, 58,* 267–288.
Villa, C., Walker, S. (2015). On the mathematics of the Jeffreys-Lindley paradox. arXiv:1503.04098.
Wang, L., Liu, X., Carroll, R. J. (2011). Estimation and variable selection for generalized additive partial linear models. *The Annals of Statistics, 39,* 1827–1851.
Wang, M., Maruyama, Y. (2016). Consistency of Bayes factor for nonnested model selection when the model dimension grows. *Bernoulli, 22*(4), 2080–2100.
Wang, M., Sun, X. (2014). Bayes factor consistency for nested linear models with a growing number of parameters. *Journal of Statistical Planning and Inference, 147,* 95–105.
Wang, X., George, E. I. (2007). Adaptive Bayesian criteria in variable selection for generalized linear models. *Statistica Sinica, 17,* 667–690.
Weisberg, S. (2005). *Applied linear regression* (3rd ed.). New York: Wiley.