Journal of Machine Learning Research 24 (2023) 1-17 Submitted 9/20; Revised 6/23; Published

Posterior Consistency for Bayesian Relevance Vector Machines

Xiao Fang XIAOFANG@UFL.EDU
Malay Ghosh GHOSHM@UFL.EDU
Department of Statistics
University of Florida
Gainesville, FL 32611, USA

Editor: John Shawe-Taylor

Abstract

Statistical modeling and inference problems with sample sizes substantially smaller than the number of available covariates are challenging. Chakraborty et al. (2012) did a full hierarchical Bayesian analysis of nonlinear regression in such situations using relevance vector machines based on reproducing kernel Hilbert space (RKHS). But they did not provide any theoretical properties associated with their procedure. The present paper revisits their problem, introduces a new class of global-local priors different from theirs, and provides results on posterior consistency as well as on posterior contraction rates.

Keywords: Global-local priors; Posterior Contraction; Reproducing kernel Hilbert space.

1. Introduction

Regression techniques are widely used virtually in any field demanding quantitative analysis. Even until today, much of this analysis relies on a linear relationship between the predictors and the response variables. This, however, is often more a convenience than reality. There is no dearth of problems of applied interest where the linearity assumption fails, and non-linear regression is called for. Fortunately, recent advancement in computer capability has allowed statisticians to tackle such non-linear regression problems. In addition, statisticians are now able to handle data where the number of covariates (say, p) far exceeds the sample size (say, n), a situation of natural ocurrence, for example in microarray experiments, image analysis, and a variety of commonly encountered problems in medicine, business, economics, sociology and others.

Chakraborty et al. (2012) considered one such problem arising from near infrared (NIR) spectroscopy where spectral measurements typically produce many more covariates (wavelets, channels) than calibration measurements (samples). They considered a full hierarchical Bayesian analysis of such data using relevance vector machines (RVM's). RVM's are machine learning techniques, originally introduced by Tipping (1999, 2001) and Bishop and Tipping (2013). These authors essentially used an empirical Bayes procedure involving Type II maximum likelihood (Good et al., 1966) estimators of prior parameters. Unlike them, Chakraborty et al. (2012) used a hierarchical Bayesian procedure by assigning distributions to the prior parameters. Hierarchical Bayes procedures typically hold advantage

over empirical Bayes procedures in that unlike the latter, they can model the uncertainty in estimating the prior parameters, thus are particularly useful for prediction.

The RVM regression approach of Chakraborty et al. (2012) was based on reproducing kernel Hilbert space (RKHS). While they could implement their procedure via Markov chain Monte Carlo (MCMC), they did not establish any theoretical properties of their method. The basic objective of this paper is to provide theoretical underpinnings to the problem introduced by Chakraborty et al. (2012). We have introduced a class of global-local priors different from the ones of Chakraborty et al. (2012). Global-local priors are widely used in high-dimensional statistics, for example, by Carvalho et al. (2010), Polson and Scott (2010) and many others. One of the attractive features of our priors is that they can handle both sparse and dense situations, and the asymptotics are based on the sample size n tending to infinity.

Our paper essentially consists of two parts. In the first part of this paper, we have proved under minimal assumptions posterior consistency as well as posterior contraction rate for a bounded kernel which includes the well-used Gaussian kernel under some mild conditions. As mentioned, the results are very general where the number of covariates can far exceed the sample size n. The prior used is a certain class of global-local priors, and the global parameter plays a key role in establishing posterior consistency as well as posterior contraction. With appropriate choice of this parameter, we are able to obtain asymptotic minimax posterior contraction rate as well. The second part of the paper deals with polynomial kernels where we are able to establish posterior consistency as well as posterior contraction rates.

The outline of the remaining sections is as follows. We have introduced the hierarchical Bayesian model in Section 2 for bounded kernels with fixed kernel parameter and have derived the marginal posterior of the regression parameter of interest. Section 3 deals with the bounded kernel and posterior consistency and contraction are established under the proposed model. Section 4 deals with results involving polynomial kernels with fixed kernel parameters. Some final remarks are made in Section 5.

2. Hierarchical Regression Model Based on RKHS

In this section we introduce the reproducing kernel Hilbert space (RKHS) and hierarchical Bayesian model based on RKHS.

2.1 Regression Model Based on RKHS

For a regression model, we have a training set $\{y_i, \boldsymbol{x}_i\}, i = 1, 2, \cdots, n$, where y_i is the response variable and $\boldsymbol{x}_i = (x_{i1}, \cdots, x_{ip})^T$ is the vector of covariates of size p corresponding to y_i. Given the training data our goal is to find an appropriate function f to predict the response y in the test set based on the covariates $\boldsymbol{x}$. This can be viewed as a regularization

problem of the form

$$\min_{f\in\mathcal{H}}[\Sigma_{i=1}^{n}L(y_i,f(\boldsymbol{x}_i))+\lambda J(f)] \tag{1}$$

where $L(y,f(\boldsymbol{x}))$ is a loss function, $J(f)$ is a penalty functional, $\lambda>0$ is the smoothing parameter, and $\mathcal{H}$ is a space of functions on which $J(f)$ is defined. In this article, we consider $\mathcal{H}$ to be a reproducing kernel Hilbert space (RKHS) with kernel K, and we denote it by $\mathcal{H}_K$. A formal definition of RKHS is given in Aronszajn (1950), Parzen (1970) and Wahba (1990).

If $f\in\mathcal{H}_K$, we take $J(f)=\parallel f\parallel_{\mathcal{H}_K}$ and rewrite (1) as

$$\min_{f\in\mathcal{H}_K}[\sum_{i=1}^{n}L(y_i,f(\boldsymbol{x}_i))+\lambda\parallel f\parallel_{\mathcal{H}_K}]. \tag{2}$$

The estimate of f is obtained as a solution of (2). It can be shown that the solution can be written as a finite sum (Wahba, 1990) and leads to a representation of f (Kimeldorf and Wahba, 1971; Wahba et al., 1999) as

$$f(\boldsymbol{x})=\Sigma_{j=1}^{n}\beta_j K(\boldsymbol{x},\boldsymbol{x}_j). \tag{3}$$

It is also a property of RKHS that

$$\parallel\sum_{j=1}^{n}\beta_j K(\boldsymbol{x},\boldsymbol{x}_j)\parallel_{\mathcal{H}_K}=\sum_{i,j=1}^{n}\beta_i\beta_j K(\boldsymbol{x}_i,\boldsymbol{x}_j).$$

To obtain the estimate of f we substitute above equation and equation (3) in (2) and then minimize it with respect to $(\beta_i,\cdots,\beta_n)$ and the smoothing parameter λ.

There are a wide variety of kernels in literature, and according to Duvenaud (2014) three basic kernels among them are the Gaussian Kernel(or squared-exponential kernel): $K(x_i,x_j)=\sigma_f^2\exp(-(x_i-x_j)^2/\theta))$, periodic kernel: $K(x_i,x_j)=\sigma_f^2\exp\left(-\frac{2}{l^2}\sin^2\left(\pi\frac{x_i-x_j}{p}\right)\right)$ and linear kernel: $K(x_i,x_j)=\sigma_f^2(x_i-c)(x_j-c)$. There are many ways to combine known kernels to get new kernels with different properties. This allows us to include as much high-level structure as necessary into our models. Two popular ways to combine kernels are addition and multiplication. A thorough discussion on the topic can be found in Duvenaud (2014), Hofmann et al. (2008), Shawe-Taylor and Cristianini (2004) and Slavakis et al. (2014).

In this paper we study the following two reproducing kernels K:
(a) The Gaussian kernel $K_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/\theta\}$, $\theta>0$,
(b) The polynomial kernel $K_\theta(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i\cdot\boldsymbol{x}_j+1)^\theta$, $\theta>0$.

We can see that the Gaussian kernel is stationary, meaning that its value only depends on the difference $\boldsymbol{x}_i-\boldsymbol{x}_j$ and the RKHS generated by the Gaussian kernel is infinite dimensional (Slavakis et al., 2014), while the polynomial kernel is nonstationary and the dimension of RKHS generated by polynomial kernel is finite.

2.2 Hierarchical Bayes Relevance Vector Machine

Begin with the model $y_i|\theta, \boldsymbol{X}_n, \boldsymbol{\beta}_n, \sigma^2 \overset{ind}{\sim} \mathcal{N}(\boldsymbol{K}_{in}^T\boldsymbol{\beta}_n, \sigma^2)$ with $\boldsymbol{K}_{in}^T = (K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_1), \cdots, K_\theta(\boldsymbol{x}_i, \boldsymbol{x}_n))$, $i = 1, \cdots, n$. We also let $\boldsymbol{Y}_n = (y_1, \cdots, y_n)^T$, $\boldsymbol{X}_n^T = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$ and $\boldsymbol{K}_n^T = (\boldsymbol{K}_{1n}, \cdots, \boldsymbol{K}_{nn})$. We call $\boldsymbol{K}_n$ the matrix associated with kernel K. In the following, we always assume the parameter θ is fixed.

The following hierarchical prior is assigned for the unknown parameters $\boldsymbol{\beta}_n$, σ^2:
Model 1:
(i) $\boldsymbol{\beta}_n|\sigma^2, \boldsymbol{\Lambda_n^2} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2\tau_n^2\boldsymbol{\Lambda}_n^2), \boldsymbol{\Lambda}_n^2 = diag(\lambda_1^2, \cdots, \lambda_n^2)$,
(ii) $\sigma^2 \sim IG(a/2, b/2)$,
(iii) $\lambda_i^2 \overset{i.i.d}{\sim} p(\lambda_i^2)$,
where $a, b > 0$ are constants not depending on n, parameter τ_n and prior distribution $p(\lambda^2)$ will be specified in the following theorems.

Remark 1 *Here we assign a global local shrinkage prior to the coefficient $\boldsymbol{\beta}_n$, and the parameter τ_n^2 is called the global shrinkage parameter. The λ_i on the other hand are local shrinkage parameters. Global local shrinkage prior is widely used in high dimensional regression problems nowadays, and it can lead to posterior consistency, see Ghosh and Chakrabarti (2017), Van Der Pas et al. (2014) and Song and Liang (2023). Our model is essentially a linear model, and $\boldsymbol{K}_n$ in our case becomes the design matrix with coefficient $\boldsymbol{\beta}_n$. We need to add some regularization conditions on $\boldsymbol{K}_n$ and also on the prior distributions of λ_i^2. Our model is similar to that of Ghosh and Chakrabarti (2017), but we assume σ^2 is unknown, which makes our analysis more complicated.*

With these priors we get

$$\begin{aligned}&\pi(\boldsymbol{\beta}_n, \sigma^2, \boldsymbol{\Lambda}_n^2|\boldsymbol{Y}_n, \boldsymbol{X}_n)\\ \propto&(\sigma^2)^{-n-a/2-1}\pi(\boldsymbol{\Lambda}_n^2)|\boldsymbol{\Lambda}_n^2|^{-1/2}\exp[\frac{-b - {\boldsymbol{\beta}_n}^T(\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})\boldsymbol{\beta}_n - (\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_n)^T(\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_n)}{2\sigma^2}];\end{aligned} \tag{4}$$

$$\boldsymbol{\beta}_n|\sigma^2, \boldsymbol{\Lambda}_n^2, \boldsymbol{Y}_n, \boldsymbol{X}_n \sim N((\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n\boldsymbol{Y}_n, \sigma^2(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}); \tag{5}$$

$$\begin{aligned}&\pi(\sigma^2, \boldsymbol{\Lambda}_n^2|\boldsymbol{Y}_n, \boldsymbol{X}_n)\\ \propto&(\sigma^2)^{-n/2-a/2-1}\pi(\boldsymbol{\Lambda}_n^2)|\boldsymbol{\Lambda}_n^2|^{-1/2}|\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}|^{-1/2}\\ &\times\exp[\frac{-b - {\boldsymbol{Y}_n}^T(\boldsymbol{I}_n - \boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n)\boldsymbol{Y}_n}{2\sigma^2}];\end{aligned} \tag{6}$$

$$\begin{aligned}&\pi(\boldsymbol{\Lambda}_n^2|\boldsymbol{Y}_n, \boldsymbol{X}_n)\\ \propto&\pi(\boldsymbol{\Lambda}_n^2)|\boldsymbol{K}_n^2\boldsymbol{\Lambda}_n^2 + \tau_n^{-2}\boldsymbol{I}_n|^{-1/2}(b + {\boldsymbol{Y}_n}^T[\boldsymbol{I}_n - \boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n]\boldsymbol{Y}_n)^{-n/2-a/2};\end{aligned} \tag{7}$$

2.3 Notations

For a vector $\boldsymbol{v} \in \mathbf{R}^n$, $\| v \|_{2,n} = (\Sigma_{i=1}^n {v_i}^2)^{\frac{1}{2}}$ denote the l_2 norm. Let $\lambda_{\max} = \max\{\lambda_1, \cdots, \lambda_n\}$, $\lambda_{\min} = \min\{\lambda_1, \cdots, \lambda_n\}$. $d_n \preccurlyeq q_n$ denotes $d_n = O(q_n)$.

3. Hierarchical Bayesian Model with Bounded Kernel

In this section, we consider the case where Model 1 has a bounded kernel with fixed parameter θ. Before studying the property of the posterior distribution, we state some regularity conditions on the matrix $\boldsymbol{K}_n$ and the true model parameters $\boldsymbol{\beta}_{0n}$, σ_0^2 .

Regularity conditions:
(A1)(Bounded Kernel) The design matrix $\boldsymbol{X}_n$ satisfies

$$c_1 \boldsymbol{I}_n \leq \boldsymbol{K}_n \leq c_2 \boldsymbol{I}_n$$

for sufficiently large n, where $c_1, c_2 > 0$ do not depend on n.

We now verify that (A1) holds for the RKHS with Gaussian kernel if the different columns $\boldsymbol{x}_i$ are sufficiently apart.

Lemma 2 *Let $\boldsymbol{K}_n$ be the matrix associate with Gaussian kernel , if $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \geq k(n) = 2\theta \log n$ for sufficiently large n for $i \neq j$, then there exists $N > 0$ such that when $n > N$, $(1 - \frac{1}{n})\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq (1 + \frac{1}{n})\boldsymbol{I}_n$.*

Proof: It suffices to show that for every $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}^T \boldsymbol{K}_n \boldsymbol{c} \leq (1 + \frac{1}{n})\boldsymbol{c}^T \boldsymbol{c}$ for large n. But

$$\begin{aligned}
&\boldsymbol{c}^T \boldsymbol{K}_n \boldsymbol{c} \\
\leq &\Sigma_{i=1}^n {c_i}^2 + \Sigma_{1 \leq i \neq j \leq n} |c_i||c_j| / \exp(k(n)/\theta) \\
= &(\Sigma_{i=1}^n {c_i}^2 + [(\Sigma_{i=1}^n |c_i|)^2 - \Sigma_{i=1}^n {c_i}^2]/(2\exp(k(n)/\theta)) \\
\leq &\left[1 - \frac{1}{2\exp(k(n)/\theta)}\right] \Sigma_{i=1}^n {c_i}^2 + \frac{n\Sigma_{i=1}^n {c_i}^2}{2\exp(k(n)/\theta)} \\
\leq &\left[1 + \frac{n-1}{2\exp(k(n)/\theta)}\right] \Sigma_{i=1}^n {c_i}^2 \\
\leq &\left(1 + \frac{1}{n}\right) \boldsymbol{c}^T \boldsymbol{c},
\end{aligned}$$

for sufficiently large n. Similarly, we have $\boldsymbol{c}^T (p^{-\theta} \boldsymbol{K}_n) \boldsymbol{c} \geq \left(1 - \frac{1}{n}\right) \boldsymbol{c}^T \boldsymbol{c}$. ■

Remark 3 *The condition that different columns $\boldsymbol{x}_i$ are sufficiently apart as $n \to \infty$ seems counter-intuitive. This condition will not hold if the data points $\boldsymbol{x}_i$ are in a compact set in $\mathbb{R}^d$, d is a fixed integer. However, the dimension p of $\boldsymbol{x}_i$ in our paper also goes to infinity as $n \to \infty$, and we do not restrict our data points on a compact set. So condition $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2 \geq k(n) = 2\theta \log n$ for sufficiently large n for $i \neq j$ is attainable, and will be satisfied if $p \gg n$.*

Remark 4 *In contrast to Ghosh and Chakrabarti (2017), who assumed $\boldsymbol{K}_n = I_n$, condition(A1) requires only boundedness of $\boldsymbol{K}_n$ in both directions.*

We now introduce the true model $\boldsymbol{Y}_n = \boldsymbol{K}_n\boldsymbol{\beta}_{0n} + \boldsymbol{\epsilon}_n$, where $\boldsymbol{\beta}_{0n} = (\beta_{01}, \cdots, \beta_{0n})^T$ and $\boldsymbol{\epsilon}_n \sim N(0, \sigma_0^2\boldsymbol{I}_n)$. We define q_n as the number of nonzero elements in $\boldsymbol{\beta}_{0n}$. Our objective is to evaluate the performance of our proposed procedure in relation to the true model. In particular, we demonstrate mean squared error consistency of the hierarchical Bayes estimator under certain regularity conditions. To this end, we first make there assumptions (A2)- (A4) in addition to (A1):
(A2) $\|\boldsymbol{\beta}_{0n}\|_2^2 = O(q_n)$,
(A3) $q_n = o(n)$,
(A4) $\sigma_0^2 = O(1)$.

Remark 5 *Condition (A2) is called sparsity assumption, it is reasonable in RVM model, because as shown in Tipping (2001), they find in practice the posterior distributions of many of the weights($\boldsymbol{\beta}_i$ in our settings) are sharply peaked around zero, and they term those training vectors associated with the remaining non-zero weights 'relevance' vectors.*

Remark 6 *Regularity condition (A3) and condition $|\beta_{0i,n}| = O(1)$ will imply regularity condition (A2), so we can also make condition $|\beta_{0i,n}| = O(1)$ as regularity condition(A2) to get all of our results in this paper, however, condition $\|\boldsymbol{\beta}_{0n}\|_2^2 = O(q_n)$ seems more natural here. Indeed, when the regression function $f(\boldsymbol{x}) = \sum_{j=1}^n \beta_j K(\boldsymbol{x}_j, \boldsymbol{x})$, the assumption $\|f\|_H \leq C_f$ can be rewritten $\beta^T K_n \beta \leq C_f$ and by Assumption (A1), we only have to assume that $\|\beta\|_2^2 \leq C_f/c_2$. This improves the connection to Gaussian processes and nonparametric estimation in RKHS.*

Theorem 7 *Assume conditions (A1)-(A4) hold. Consider the priors assigned to $\boldsymbol{\Lambda}_n$ and σ^2 in Section 2.2. Then if $\tau_n^2 = O(n^{-2})$ and*

$$\int \lambda^4 p(\lambda^2) d\lambda^2 < \infty,$$

$$\mathbb{E}_0 \parallel \mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n, \boldsymbol{X}_n) - \boldsymbol{K}_n\boldsymbol{\beta}_{0n} \parallel_{2,n}^2 = o(n)$$

as $n \to \infty$, where $\mathbb{E}_0$ denotes expectation under the true model.

Remark 8 *This theorem holds for both $p \leq n$ and $p > n$.*

Proof of Theorem 7: First we calculate

$$
\begin{aligned}
&\mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\\
=&\mathbb{E}(\mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n^2,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n)-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\\
=&\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n\boldsymbol{Y}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n])-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\\
=&\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&+\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\\
=&\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&+\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}-\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\\
=&\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&-\mathbb{E}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
=&I-II,
\end{aligned}
\tag{8}
$$

where we let the first term in the second last equality in (8) be I and the second term be II.

In view of (8), it suffices to show that $\mathbb{E}_0 \parallel I \parallel_{2,n}^2 = o(n)$ and $\mathbb{E}_0 \parallel II \parallel_{2,n}^2 = o(n)$.

In order to prove these results, first we recall the matrix result that if $\boldsymbol{A} \geq \boldsymbol{B}$, that is $\boldsymbol{A}-\boldsymbol{B}$ is nonnegative definite, then $\boldsymbol{CAC}^T \geq \boldsymbol{CBC}^T$.

Next recalling (A1),

$$
\begin{aligned}
&\parallel I \parallel_{2,n}^2\\
\leq&\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&c_2^2\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-2}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
=&c_2^2\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\\
&\cdot(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&c_2^2\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}(\boldsymbol{K}_n^2)^{-1}\\
&\cdot(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&(c_2^2/c_1^2)\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&(c_2^2/c_1^2)\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T(\boldsymbol{K}_n\tau_n^2\boldsymbol{\Lambda}_n^2\boldsymbol{K}_n)(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&\tau_n^2(c_2^2/c_1^2)\mathbb{E}[(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})^T\lambda_{\max}^2\boldsymbol{K}_n^2(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\leq&\frac{c_2^4\tau_n^2}{c_1^2}\mathbb{E}[\lambda_{\max}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n]\cdot\|\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2.
\end{aligned}
\tag{9}
$$

Next applying the Cauchy-Schwarz inequality and the fact that $\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2 \sim \sigma_0^2\chi_n^2$ under the true model $\boldsymbol{Y}_n \sim \mathcal{N}(\boldsymbol{K}_n\boldsymbol{\beta}_{0n}, \sigma_0^2\boldsymbol{I}_n)$,

$$\begin{aligned}
&\mathbb{E}_0[\mathbb{E}[\lambda_{max}^2|\boldsymbol{Y}_n, \boldsymbol{X}_n] \cdot \|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2] \\
\leq &\mathbb{E}_0^{1/2}(\mathbb{E}^2[\lambda_{max}^2|\boldsymbol{Y}_n, \boldsymbol{X}_n])\mathbb{E}_0^{1/2}[\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^4] \\
\leq &\mathbb{E}_0^{1/2}(\mathbb{E}[\lambda_{max}^4|\boldsymbol{Y}_n, \boldsymbol{X}_n]) \cdot \mathbb{E}_0^{1/2}(\sigma_0^2\chi_n^2)^2 \\
= &\mathbb{E}_0^{1/2}(\lambda_{max}^4) \cdot \sqrt{n(n+2)\sigma_0^4} \\
\leq &\mathbb{E}_0^{1/2}(\sum_{i=1}^n \lambda_i^4) \cdot \sqrt{n(n+2)\sigma_0^4} \\
= &(\sum_{i=1}^n \mathbb{E}_0\lambda_i^4)^{1/2} \cdot \sqrt{n(n+2)\sigma_0^4} \\
= &\sqrt{n}\mathbb{E}_0^{1/2}(\lambda_1^4) \cdot \sqrt{n(n+2)\sigma_0^4},
\end{aligned} \tag{10}$$

The result now follows from (9) and (10) since $\tau_n^2 = O(n^{-2})$.

Next using $c_1^2\boldsymbol{I}_n \leq \boldsymbol{K}_n^2 \leq c_2^2\boldsymbol{I}_n$ and assumption (A2),

$$\begin{aligned}
&\| \, II \, \|_{2,n}^2 \\
\leq &\mathbb{E}[\|\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}\|_{2,n}^2|\boldsymbol{Y}_n, \boldsymbol{X}_n] \\
= &\tau_n^{-4}\mathbb{E}[\boldsymbol{\beta}_{0n}^T\boldsymbol{\Lambda}_n^{-2}(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n, \boldsymbol{X}_n] \\
\leq &c_2^2\tau_n^{-4}\mathbb{E}[\boldsymbol{\beta}_{0n}^T\boldsymbol{\Lambda}_n^{-2}(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n, \boldsymbol{X}_n] \\
\leq &c_2^2\tau_n^{-4}\mathbb{E}[\boldsymbol{\beta}_{0n}^T\boldsymbol{\Lambda}_n^{-2}(c_1^2\boldsymbol{I}_n + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n, \boldsymbol{X}_n] \\
\leq &c_2^2\tau_n^{-4}\mathbb{E}[\boldsymbol{\beta}_{0n}^T\boldsymbol{\Lambda}_n^{-2}\tau_n^4\boldsymbol{\Lambda}_n^4\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n, \boldsymbol{X}_n] \\
= &c_2^2\|\boldsymbol{\beta}_{0n}\|_{2,n}^2 = O(q_n) = o(n).
\end{aligned} \tag{11}$$

This completes the proof of the theorem. ■

Remark 9 *The assumption of finiteness of the second moment of λ^2 can be weakened. All we need is the finiteness of the $(1+\delta)$th moment of λ^2, where $\delta > 0$. To see this, one applies Holder's inequality to get*

$$\begin{aligned}
&\mathbb{E}_0[\mathbb{E}[\lambda_{\max}^2|\boldsymbol{Y}_n, \boldsymbol{X}_n] \cdot \|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2] \\
\leq &\mathbb{E}_0^{\frac{1}{1+\delta}}(\mathbb{E}^{1+\delta}[\lambda_{\max}^2|\boldsymbol{Y}_n, \boldsymbol{X}_n])\mathbb{E}_0^{\frac{\delta}{1+\delta}}[\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^{2\cdot\frac{1+\delta}{\delta}}] \\
\leq &\mathbb{E}_0^{\frac{1}{1+\delta}}(\mathbb{E}[\lambda_{\max}^{2\cdot(1+\delta)}|\boldsymbol{Y}_n, \boldsymbol{X}_n])\mathbb{E}_0^{\frac{\delta}{1+\delta}}[\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^{2\cdot\frac{1+\delta}{\delta}}] \\
= &\mathbb{E}_0^{\frac{1}{1+\delta}}(\lambda_{\max}^{2\cdot(1+\delta)})\mathbb{E}_0^{\frac{\delta}{1+\delta}}[\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^{2\cdot\frac{1+\delta}{\delta}}].
\end{aligned} \tag{12}$$

Since $\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2 \sim \sigma_0^2\chi_n^2$,

$$\mathbb{E}_0[\|\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^{2\cdot\frac{1+\delta}{\delta}}] = (\sigma_0^2)^{\frac{1+\delta}{\delta}}(\chi_n^2)^{1+1/\delta} = (2\sigma_0^2)^{\frac{1+\delta}{\delta}}\Gamma(n/2 + 1/\delta + 1)/\Gamma(n/2).$$

Using Stirling's formula, $\Gamma(n/2+1/\delta+1)/\Gamma(n/2) \leq Cn^{1/\delta+1}$*, so that the second term in the right hand side of (12) is bounded above by a constant multiple of* n*, also*

$$\mathbb{E}(\lambda_{\max}^{2\cdot(1+\delta)})\mathbb{E}^{\frac{1}{1+\delta}}(\lambda_{\max}^{2\cdot(1+\delta)}) \leq n^{\frac{1}{1+\delta}}\mathbb{E}^{\frac{1}{1+\delta}}\left(\lambda_1^{2\cdot(1+\delta)}\right). \tag{13}$$

By (12) and (13),

$$\mathbb{E}_0[\mathbb{E}[\lambda_{\max}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n]\cdot\|\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2] \leq Cn^{1+\frac{1}{1+\delta}}\mathbb{E}^{\frac{1}{1+\delta}}\left(\lambda_1^{2\cdot(1+\delta)}\right). \tag{14}$$

Then Theorem 1 still holds since $\tau_n^2 = O(n^{-2})$*.*

Remark 10 *The assumption of* $(1+\delta)$*th moment of* λ^2 *holds for several distributions. Examples include the common Gamma distribution, the inverse Gaussian distribution, Student's t-distribution with finite second moment, the inverse gamma distribution with shape parameter greater than* $1+\delta$ *and the beta prime priors* $p(\lambda^2) \propto (\lambda^2)^{a-1}(1+\lambda^2)^{-a-b}$ *with* $b > 1+\delta$*.*

Remark 11 *Checking the proof of Theorem 7, we can reformulate it as follows: Assume conditions (A1)-(A4) hold. Consider the priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *in Section 2.2. Then if* $\tau_n^2 \preccurlyeq n^{-3/2}q_n$ *and*

$$\int \lambda^4 p(\lambda^2)d\lambda^2 < \infty,$$

$$\mathbb{E}_0 \parallel \mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n) - \boldsymbol{K}_n\boldsymbol{\beta}_{0n} \parallel_{2,n}^2 \preccurlyeq q_n$$

as $n \to \infty$*.*

Theorem 12 *Assume conditions (A1)-(A4) hold. Consider the priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *as in Section 2.3. Then if* $\tau_n^2 = O(n^{-2})$ *and*

$$\int \lambda^4 p(\lambda^2)d\lambda^2 < \infty,$$

then

$$\mathbb{E}_0\{tr[\mathbb{V}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\} = o(n)$$

as $n \to \infty$*.*

Proof of Theorem 12: By (6) we have $\mathbb{E}(\sigma^2|\boldsymbol{\Lambda}_n,\boldsymbol{Y}_n,\boldsymbol{X}_n) = \frac{b+\boldsymbol{Y}_n{}^T(\boldsymbol{I}_n-\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n)\boldsymbol{Y}_n}{n+a-2}$,

$$\begin{aligned}
&tr[\mathbb{V}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\\
=&tr\mathbb{E}[\mathbb{V}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&+tr\mathbb{V}[\mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
=&tr\mathbb{E}[\frac{b+\boldsymbol{Y}_n{}^T(\boldsymbol{I}_n-\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n)\boldsymbol{Y}_n}{n+a-2}\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&+tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n\boldsymbol{Y}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
=&III+IV,
\end{aligned} \tag{15}$$

where we let the first term in the second last equality in (15) be III and the second term be IV.

Since $\boldsymbol{K}_n^2 \leq c_2^2 \boldsymbol{I}_n$, $tr({\boldsymbol{K}_n}^2) \leq nc_2^2$. Hence, by Cauchy-Schwartz inequality,

$$\begin{aligned} III \leq & tr\mathbb{E}[\frac{b+{\boldsymbol{Y}_n}^T\boldsymbol{Y}_n}{n+a-2}(\tau_n^2\lambda_{\max}^2\boldsymbol{K}_n^2)|\boldsymbol{Y}_n,\boldsymbol{X}_n] \leq nc_2^2\tau_n^2\mathbb{E}[\lambda_{\max}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n]\frac{b+{\boldsymbol{Y}_n}^T\boldsymbol{Y}_n}{n+a-2} \\ \leq & nc_2^2\tau_n^2\mathbb{E}^{1/2}(\lambda_{\max}^4|\boldsymbol{Y}_n,\boldsymbol{X}_n)\frac{\mathbb{E}^{1/2}(b+{\boldsymbol{Y}_n}^T\boldsymbol{Y}_n)^2}{n+a-2} \end{aligned} \tag{16}$$

We may note that under the true model $\boldsymbol{Y}_n \sim \mathcal{N}(\boldsymbol{K}_n\boldsymbol{\beta}_{0n}, \sigma_0^2\boldsymbol{I}_n)$, $\boldsymbol{Y}_n^T\boldsymbol{Y}_n/\sigma_0^2$ is a noncentral chisquared distribution with degrees of freedom n and noncentral parameter $(\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n})/\sigma_0^2$. Hence, $\mathbb{E}(\boldsymbol{Y}_n^T\boldsymbol{Y}_n/\sigma_0^2) = n + (\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n})/\sigma_0^2$ and $\mathbb{V}(\boldsymbol{Y}_n^T\boldsymbol{Y}_n/\sigma_0^2) = 2n + 8(\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n})/\sigma_0^2$. Thus

$$\begin{aligned} \mathbb{E}_0^{1/2}(b+{\boldsymbol{Y}_n}^T\boldsymbol{Y}_n)^2 \leq & \mathbb{E}_0^{1/2}[2b^2 + 2({\boldsymbol{Y}_n}^T\boldsymbol{Y}_n)^2] \leq \sqrt{2b^2} + \mathbb{E}_0^{1/2}[2({\boldsymbol{Y}_n}^T\boldsymbol{Y}_n)^2] \\ = & \sqrt{2}b + \sqrt{2}[2n\sigma_0^2 + 8\sigma_0^2(\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}) + (n\sigma_0^2 + (\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}))^2]^{1/2} \\ \leq & \sqrt{2}b + \sqrt{2}[2n\sigma_0^2 + 8c_2^2q_n + (n\sigma_0^2 + c_2^2q_n)^2]^{1/2} = O(n) \end{aligned} \tag{17}$$

Combining (16) and (17)

$$\begin{aligned} \mathbb{E}_0 III \leq & nc_2^2\tau_n^2\mathbb{E}_0[n\mathbb{E}(\lambda_1^4)]^{1/2}O(n)/(n+a-2) \\ \preccurlyeq & c_2^2n^{3/2}\tau_n^2 = o(n). \end{aligned}$$

$$\begin{aligned} IV = & tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n} + \boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n] \\ \leq & 2tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n] \\ & + 2tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]. \end{aligned} \tag{18}$$

Now noting that $tr[\mathbb{V}(X)] \leq E\|X\|^2$, the 1st term in the RHS of (18)

$$\leq 2\mathbb{E}[\|\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n})\|_{2,n}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n].$$

Then by (9), (10),we have

$$\mathbb{E}_0\{tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\} \preccurlyeq q_n = o(n). \tag{19}$$

The 2nd term in the RHS of (18)

$$\begin{aligned} \leq & 2\mathbb{E}[\|\boldsymbol{K}_n(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\|_{2,n}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n] \\ \leq & 2c_2^2\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-2}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n} \\ = & 2c_2^2\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n} \\ \leq & 2(c_2^2/c_1^2)\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2 + \tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n} \\ \leq & 2(c_2^2/c_1^2)\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2)^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n} \\ \leq & 2\frac{c_2^4}{c_1^2}\|\boldsymbol{\beta}_{0n}\|_{2,n}^2 = O(q_n) = o(n). \end{aligned}$$

Hence,

$$\mathbb{E}_0\{tr\mathbb{V}[\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\}=o(n). \tag{20}$$

The theorem follows from (15)-(20). ■

Remark 13 *Checking the proof of Theorem 12, we can reformulate it as follows: Assume conditions (A1)-(A4) hold. Consider the priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *as in Section 2.3. Then if* $\tau_n^2 \preccurlyeq n^{-3/2}q_n$ *and*

$$\int \lambda^4 p(\lambda^2)d\lambda^2 < \infty,$$

$$\mathbb{E}_0\{tr[\mathbb{V}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\} \preccurlyeq q_n$$

as $n \to \infty$.

Combining Remark 11 and Remark 13, we can get:

Corollary 14 *Assume conditions (A1)-(A4) hold. Then with the prior assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *in model 1, if* $\tau_n^2 \preccurlyeq n^{-\frac{3}{2}}q_n$ *and*

$$\int \lambda^4 p(\lambda^2)d\lambda^2 < \infty,$$

$$\mathbb{E}_0 P(\|\boldsymbol{K}_n\boldsymbol{\beta}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|_{2,n}^2 \geq M_n q_n|\boldsymbol{Y}_n,\boldsymbol{X}_n) \to 0$$

as $n \to \infty$, *where* $M_n \to \infty$ *as* $\to \infty$.

In particular, one may take $M_n = \log(n/q_n)$ to get the asymptotic minimax contraction bound.

4. Hierarchical Bayesian Model with Polynomial Kernel

For a polynomial kernel, we can not apply Theorem 7 directly, because the regularity condition (A1) does not generally hold. For example, for polynomial kernel with fixed parameter $\theta > 0$, if the design matrix $\boldsymbol{X}_n$ satisfies the orthogonality condition, namely, $\boldsymbol{X}_n\boldsymbol{X}_n^T = p\boldsymbol{I}_n$, then $p^{-\theta}\boldsymbol{K}_n - I_n \to 0$ as $n \to \infty$ so that $\frac{1}{2}p^\theta\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq 2p^\theta\boldsymbol{I}_n$ for sufficiently large n, which does not satisfy condition (A1). In this section, we consider the case when the design matrix $\boldsymbol{X}_n$ satisfies $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$, where $t_1(n)$ and $t_2(n)$ are functions depending solely on n.

We have the following posterior contraction result for polynomial kernels.

Theorem 15 *Assume conditions (A2)-(A4) hold. Then with the prior assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *in Model 1, if* $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$, $t_2^2(n) \preccurlyeq \log(\frac{n}{q_n})$, $\tau_n^2 \preccurlyeq n^{-3/2}q_n$ *and*

$$\int \lambda^4 p(\lambda^2)d\lambda^2 < \infty,$$

$$\mathbb{E}_0 \parallel \mathbb{E}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n) - \boldsymbol{K}_n\boldsymbol{\beta}_{0n} \parallel^2_{2,n} \preccurlyeq q_n \log(\frac{n}{q_n})$$

as $n \to \infty$.

Proof of Theorem 15: This proof is almost the same as of Theorem 7. It suffices to show that $\mathbb{E}_0 \parallel I \parallel^2_{2,n} \preccurlyeq q_n \log(\frac{n}{q_n})$ and $\mathbb{E}_0 \parallel II \parallel^2_{2,n} \preccurlyeq q_n \log(\frac{n}{q_n})$ as $n \to \infty$.

Substituting $t_1(n)$ and $t_2(n)$ for c_1 and c_2 in Theorem 7, we get

$$\mathbb{E}_0 \parallel I \parallel^2_{2,n} \leq \sqrt{n}\frac{t_2^4(n)\tau_n^2}{t_1^2(n)}\mathbb{E}_0^{1/2}(\lambda_1^4) \cdot \sqrt{n(n+2)\sigma_0^4} \preccurlyeq n^{3/2}\tau_n^2 t_2^2(n) \preccurlyeq q_n \log(\frac{n}{q_n}) \tag{21}$$

and

$$\mathbb{E}_0 \parallel II \parallel^2_{2,n} \preccurlyeq t_2^2(n)\|\boldsymbol{\beta}_{0n}\|^2_{2,n} \preccurlyeq q_n \log(\frac{n}{q_n}). \tag{22}$$

■

Theorem 16 *Assume conditions (A2)-(A4) hold. Then with the same priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *in Section 2.2, if* $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$, $t_2^2(n) \preccurlyeq \log(\frac{n}{q_n})$, $\tau_n^2 \preccurlyeq n^{-3/2}q_n$ *and*

$$\int \lambda^4 p(\lambda^2) d\lambda^2 < \infty,$$

then

$$\mathbb{E}_0\{tr[\mathbb{V}(\boldsymbol{K}_n\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\} \preccurlyeq q_n \log(\frac{n}{q_n})$$

as $n \to \infty$.

Proof of Theorem 16: This proof is almost the same as of Theorem 12. Substitute c_1, c_2 in the proof of Theorem 12 by $t_1(n)$, $t_2(n)$. ■

Combining Theorems 15 and 16 we get minimax contraction rate for our model.

Corollary 17 *Assume conditions (A2)-(A4) hold. Then with the same priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *in Section 2.2, if* $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$, $t_2^2(n) \preccurlyeq \log(\frac{n}{q_n})$, $\tau_n^2 \preccurlyeq n^{-3/2}q_n$ *and*

$$\int \lambda^4 p(\lambda^2) d\lambda^2 < \infty,$$

$$\mathbb{E}_0 P(\|\boldsymbol{K}_n\boldsymbol{\beta}_n - \boldsymbol{K}_n\boldsymbol{\beta}_{0n}\|^2_{2,n} \geq q_n \log(\frac{n}{q_n})|\boldsymbol{Y}_n,\boldsymbol{X}_n) \to 0$$

as $n \to \infty$.

Remark 18 *For polynomial kernels with fixed parameters, if the design matrix is approximately orthogonal , then we still have* $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$. *Theorems 15 , 16 and Corollary 17 hold if* $p^\theta \preccurlyeq \sqrt{q_n \log(\frac{n}{q_n})}$, *then* $p < n$. *Actually we have the following lemma which describes the behavior of the kernel* K *when the design matrix is approximately orthogonal.*

Lemma 19 *Let* $\boldsymbol{K}$ *be a polynomial kernel with parameter* $\theta \in [a_L, a_U]$, $a_L > 1/2$, *and the design matrix* $\boldsymbol{X}_n$ *satisfies*

$$|\frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_i + 1}{p} - 1| \leq \frac{1}{h(n)} \quad \textit{and}$$
$$|\frac{\boldsymbol{x}_i \cdot \boldsymbol{x}_j + 1}{p}| \leq \frac{1}{k(n)}, 1 \leq i \neq j \leq n,$$

$h(n) = 2a_U \cdot n$, $k(n) = n^4$. *Then for sufficiently large* n, *there exists* $N > 0$ *such that when* $n > N$, $(1 - \frac{1}{n})\boldsymbol{I}_n \leq p^{-\theta}\boldsymbol{K}_n \leq (1 + \frac{1}{n})\boldsymbol{I}_n$ *for all* $\theta \in [a_L, a_U]$.

Proof: *It suffices to show that for every* $\boldsymbol{c} \neq \boldsymbol{0}$, $\boldsymbol{c}^T(p^{-\theta}\boldsymbol{K}_n)\boldsymbol{c} \leq (1 + \frac{1}{n})\boldsymbol{c}^T\boldsymbol{c}$ *for large* n. *But*

$$\begin{aligned}
&\boldsymbol{c}^T \left(p^{-\theta}\boldsymbol{K}_n\right) \boldsymbol{c} \\
\leq &\Sigma_{i=1}^n {c_i}^2 \left(1 + \frac{1}{h(n)}\right)^\theta + \Sigma_{1 \leq i \neq j \leq n} |c_i||c_j|/k^\theta(n) \\
= &\left(1 + \frac{1}{h(n)}\right)^\theta \Sigma_{i=1}^n {c_i}^2 + [(\Sigma_{i=1}^2 |c_i|)^2 - \Sigma_{i=1}^2 {c_i}^2]/(2k^\theta(n)) \\
\leq &\left[\left(1 + \frac{1}{h(n)}\right)^\theta - \frac{1}{2k^\theta(n)}\right] \Sigma_{i=1}^n {c_i}^2 + \frac{n\Sigma_{i=1}^2 {c_i}^2}{2k^\theta(n)} \\
\leq &\left[\left(1 + \frac{1}{h(n)}\right)^\theta - \frac{1}{2n^{4\theta}} + \frac{1}{2n^{4\theta-1}}\right] \Sigma_{i=1}^n {c_i}^2 \\
\leq &\left(1 + \frac{1}{n}\right) \boldsymbol{c}^T\boldsymbol{c}, \ \textit{(since } \theta > 1/2\textit{)}.
\end{aligned}$$

Similarly, we have $\boldsymbol{c}^T(p^{-\theta}\boldsymbol{K}_n)\boldsymbol{c} \geq (1 - \frac{1}{n})\boldsymbol{c}^T\boldsymbol{c}$. ■

Although we have to assume $p < n$ to get posterior contraction results for polynomial kernels, we can still get posterior consistency for the $p > n$ case for polynomial kernels.

Theorem 20 *Assume conditions (A2)-(A4) hold. Consider the same priors assigned to* $\boldsymbol{\Lambda}_n$ *and* σ^2 *as in Section 2.2, If* $t_1(n)\boldsymbol{I}_n \leq \boldsymbol{K}_n \leq t_2(n)\boldsymbol{I}_n$, $t_2(n)/t_1(n) = O(1)$ *and* $t_1(n) \succ 1$, $\tau_n^2 \preccurlyeq n^{-3/2}q_n$,

$$\int \lambda^4 p(\lambda^2) d\lambda^2 < \infty, \qquad \int \lambda^{-2} p(\lambda^2) d\lambda^2 < \infty,$$

then

$$\mathbb{E}_0 P(\|\boldsymbol{\beta}_n - \boldsymbol{\beta}_{0n}\|_{2,n}^2 \geq M_n q_n | \boldsymbol{Y}_n, \boldsymbol{X}_n) \to 0$$

as $n \to \infty$, *where* $M_n \to \infty$ *as* $n \to \infty$.

Remark 21 *For polynomial kernels, condition $t_1(n) \succ 1$ implies $p \succ 1$, that is $p \to \infty$ as $n \to \infty$.*

Proof of Theorem 20: First calculate

$$\begin{aligned}\mathbb{E}(\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)-\boldsymbol{\beta}_{0n} =&\mathbb{E}(\mathbb{E}(\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n^2,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n)-\boldsymbol{\beta}_{0n}\\ =&\mathbb{E}((\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n\boldsymbol{Y}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)\\ &-\mathbb{E}[(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})\boldsymbol{\beta}_{0n})]\\ =&\mathbb{E}[(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n(\boldsymbol{Y}_n-\boldsymbol{K}_n\boldsymbol{\beta}_{0n})|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ &-\mathbb{E}[(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ =&V_1-V_2.\end{aligned} \tag{23}$$

where we let the first term in the second last equality in (23) be V_1 and the second term be V_2.

First, we show that $\mathbb{E}_0 \parallel V_1 \parallel_{2,n}^2 \preccurlyeq q_n$. Similar to (9) and (10) in the proof of Theorem 7, we get

$$\mathbb{E}_0 \parallel V_1 \parallel_{2,n}^2 \leq \sqrt{n}\frac{t_2^2(n)\tau_n^2}{t_1^2(n)}\mathbb{E}_0^{1/2}(\lambda_1^4)\cdot\sqrt{n(n+2)\sigma_0^4} \preccurlyeq n^{3/2}\tau_n^2 \preccurlyeq q_n. \tag{24}$$

Next we show $\mathbb{E}_0 \parallel V_2 \parallel_{2,n}^2 \preccurlyeq q_n$, since $\boldsymbol{K}_n^4 \geq t_1^4(n)\boldsymbol{I}_n$,

$$\begin{aligned}&\mathbb{E}_0 \parallel V_2 \parallel_{2,n}^2\\ \leq&\mathbb{E}_0\mathbb{E}[\|(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2}\boldsymbol{\beta}_{0n}\|_{2,n}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ \leq&t_1^{-4}(n)\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T(\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^4(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}(\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ =&t_1^{-4}(n)\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ =&t_1^{-4}(n)\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1/2}(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1}(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1/2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ \leq&t_1^{-4}(n)\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1/2}\boldsymbol{K}_n^2(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1/2}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ \leq&(t_2^2(n)/t_1^4(n))\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T(\boldsymbol{K}_n^{-2}+\tau_n^2\boldsymbol{\Lambda}_n^2)^{-1}\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\ \leq&(t_2^2(n)/t_1^4(n))\mathbb{E}_0\mathbb{E}[\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n] \leq (t_2^4(n)/t_1^4(n))\|\boldsymbol{\beta}_{0n}\|_{2,n}^2 \preceq q_n.\end{aligned} \tag{25}$$

Then we have

$$\mathbb{E}_0 \parallel \mathbb{E}(\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)-\boldsymbol{\beta}_{0n} \parallel_{2,n}^2 \preccurlyeq q_n.$$

as $n \to \infty$.

Now we will show

$$\mathbb{E}_0\{tr[\mathbb{V}(\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\} \preccurlyeq q_n,$$

as $n \to \infty$.

The method of proof is similar to that in Theorem 12. We only provide some details of the key steps here.

$$\begin{aligned}
&tr[\mathbb{V}(\boldsymbol{\beta}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n)]\\
=&tr\mathbb{E}[\mathbb{V}(\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n]+tr\mathbb{V}[\mathbb{E}(\boldsymbol{\beta}_n|\sigma^2,\boldsymbol{\Lambda}_n,\boldsymbol{Y}_n,\boldsymbol{X}_n)|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
=&tr\mathbb{E}[\frac{b+{\boldsymbol{Y}_n}^T(\boldsymbol{I}_n-\boldsymbol{K}_n(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n)\boldsymbol{Y}_n}{n+a-2}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
&+tr\mathbb{V}[(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n\boldsymbol{Y}_n|\boldsymbol{Y}_n,\boldsymbol{X}_n]
\end{aligned}\tag{26}$$

$$\begin{aligned}
&tr\mathbb{V}[(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\le&2\mathbb{E}[\|(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\|_{2,n}^2|\boldsymbol{Y}_n,\boldsymbol{X}_n]\\
\le&2\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-2}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\\
=&2\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1/2}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\\
\le&2(1/t_1^2(n))\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2+\tau_n^{-2}\boldsymbol{\Lambda}_n^{-2})^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\\
\le&2(1/t_1^2(n))\boldsymbol{\beta}_{0n}^T\boldsymbol{K}_n^2(\boldsymbol{K}_n^2)^{-1}\boldsymbol{K}_n^2\boldsymbol{\beta}_{0n}\\
\le&2\frac{t_2^2(n)}{t_1^2(n)}\|\boldsymbol{\beta}_{0n}\|_{2,n}^2=O(q_n)=o(n).
\end{aligned}\tag{27}$$

Combining above equations, we prove this theorem. ■

5. Discussion

Tipping (2001), Williams and Rasmussen (2006) pointed out that RVM is a special case of Gaussian process. van der Vaart and van Zanten (2008) and Ghosal and Van der Vaart (2017) obtained several posterior concentration results for Gaussian process models. They considered estimating a regression function f based on observations $y_1,\cdots,y_n$ in a normal regression model with fixed covariates $y_i=f(\boldsymbol{x}_i)+\epsilon_i$, where $\epsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0,\sigma_0^2)$ and the covariates $\boldsymbol{x}_1,\cdots,\boldsymbol{x}_n$ are fixed elements from a set $\mathcal{X}$.

A prior on f is induced by setting $f(\boldsymbol{x})=W_{\boldsymbol{x}}$ for a Gaussian process $(W_{\boldsymbol{x}}: x\in\mathcal{X})$. Any Gaussian element in a separable Banach space can be expanded as an infinite series $\sum_i Z_i h_i$ for i.i.d standard normal variables Z_i and elements h_i from its RKHS. van der Vaart and van Zanten (2008) truncated this infinite series at a sufficient high level to get a new Gaussian process prior. If this series converges to the infinite series quickly, then by Theorem 2.2 in van der Vaart and van Zanten (2008), the same posterior rate of contraction is attained. Since finite sums may be easier to handle, it is interesting to investigate special expansions and the number of terms that need to be retained in order to obtain the same contraction rate. van der Vaart and van Zanten (2008) illustrated this by an example of the truncated wavelet expansion of functions in $\mathbb{L}_2([0,1]^d)$. Ghosal and Van Der Vaart (2007) considered the truncated B-spline expansion in their Theorem 12. These truncated series are quite similar to our model if set $p=d$ fixed, $\boldsymbol{K}_n=\boldsymbol{I}_n$ and the prior $\boldsymbol{\beta}_n\sim\mathcal{N}(\boldsymbol{0},\boldsymbol{I}_n)$. However in their case, the number of terms in the random series is $O(n^\alpha)$, $\alpha<1$, while ours

is n. Hence, adding global shrinkage parameter τ_n to accommodate sparsity seems reasonable, meanwhile we also introduce hierarchical model to make the prior of $\boldsymbol{\beta}_n$ have heavy tail, which helps to detect nonzero coefficients. Also, the Gaussian process prior related to RVM is data dependent (Williams and Rasmussen, 2006), which is likely to add flexibility to prediction.

In our model, we assume the parameters in kernel K is fixed, however, in Tipping (2001), they argue that for Gaussian kernel, the data set can become more probable at some intermediate width(θ in our paper), so we could put priors on kernel parameters as in Chakraborty et al. (2012), and its posterior contraction properties remain to be explored.

Acknowledgments

We would like to thank the referees for their many valuable comments which improved the paper considerably.

References

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Christopher M Bishop and Michael Tipping. Variational relevance vector machines. *arXiv preprint arXiv:1301.3838*, 2013.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Sounak Chakraborty, Malay Ghosh, and Bani K Mallick. Bayesian nonlinear regression for large p small n problems. *Journal of Multivariate Analysis*, 108:28–40, 2012.

David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

Prasenjit Ghosh and Arijit Chakrabarti. Asymptotic optimality of one-group shrinkage priors in sparse high-dimensional problems. *Bayesian Analysis*, 12(4):1133–1161, 2017.

Irving John Good, Ian Hacking, RC Jeffrey, and Håkan Törnebohm. The estimation of probabilities: An essay on modern bayesian methods. *Synthese*, 16(2), 1966.

Thomas Hofmann, Bernhard Schölkopf, and Alexander J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220, 2008.

George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

Emanuel Parzen. Statistical inferences on time series by RKHS methods. *Proceedings of the 12th Biennial Seminar, Canadian Mathematical Congress, Montreal, Canada*, pages 1–37, 1970.

Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse bayesian regularization and prediction. *Bayesian statistics*, 9(501-538):105, 2010.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Konstantinos Slavakis, Pantelis Bouboulis, and Sergios Theodoridis. Online learning in reproducing kernel hilbert spaces. In *Academic Press Library in Signal Processing*, volume 1, pages 883–987. Elsevier, 2014.

Qifan Song and Faming Liang. Nearly optimal bayesian shrinkage for high-dimensional regression. *Science China Mathematics*, 66(2):409–442, 2023.

Michael Tipping. The relevance vector machine. *Advances in neural information processing systems*, 12, 1999.

Michael Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Stéphanie L Van Der Pas, Bas JK Kleijn, and Aad W Van Der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *The Annals of Statistics*, 36(3):1435 – 1463, 2008.

Grace Wahba. *Spline models for observational data*. SIAM, 1990.

Grace Wahba et al. Support vector machines, reproducing kernel hilbert spaces and the randomized gacv. *Advances in Kernel Methods-Support Vector Learning*, 6:69–87, 1999.

Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.