Supplementary Materials for "Functional Horseshoe Priors for Subspace Shrinkage"

A A description of the B-spline Basis Functions

The B-splines basis functions can be constructed in a recursive manner. Let the positive integer q denote the degree of the B-spline basis functions satisfying $k_n > q + 1$. Define a sequence of knots $0 = t_0 < t_1 < \cdots < t_{k_n - q} = 1$. In addition, define q knots $t_{-q} = \cdots = t_{-1} = t_0$ and another set of q knots $t_{k_n - q} = \cdots = t_{k_n}$. As in De Boor (2001), the B-spline basis functions are defined as

$$\begin{aligned}
\phi_{j,1}(x) &= \begin{cases} 1, & t_j \le x < t_{j+1}, \\ 0, & \text{otherwise}, \end{cases} \\
\phi_{j,q+1}(x) &= \frac{x - t_j}{t_{j+q} - t_j}\phi_{j,q}(x) + \frac{t_{j+q+1} - x}{t_{j+q+1} - t_{j+1}}\phi_{j+1,q}(x),
\end{aligned}$$

for $j = -q, \ldots, k_n - q - 1$. We reindex $j = -q, \ldots, k_n - q - 1$ to $j = 1, \ldots, k_n$.

We state a set of standard regularity conditions that have been used by others (Zhou et al. (1998), Claeskens et al. (2009)) to prove minimax optimality of B-spline estimators. We assume that:

(A1). Let $u = \max_{1 \le j \le (k_n - 1)}(t_{j+1} - t_j)$. There exists a constant $C > 0$, such that $u \le C \min_{1 \le j \le (k_n - 1)}(t_{j+1} - t_j)$ and $u = o(k_n^{-1})$.

(A2). There exists a distribution function G with a positive continuous density function that satisfies $\sup_{x \in [0,1]} |G_n(x) - G(x)| = o(k_n^{-1})$, where G_n is the empirical distribution of the covariates $\{x_i\}_{1 \le i \le n}$, which are assumed to be fixed by design.

B Additional Simulation Studies for Univariate Examples

In this section, we provide details for the replicated studies for the varying coefficient model (ii) and log-density model (iii) in (11) in the main article, that were skipped for space constraints in Section 4 of the main article.

For (ii), we generated the covariates independently from a uniform distribution between $-\pi$ and π and set the error variance $\sigma^2 = 1$. For each case (ii) and (iii), we considered three parametric choices for f. For case (ii), we considered constant, quadratic and sinusoidal functions. For (iii), we considered normal, log-normal and mixture of normal distributions. For the first two cases, we standardized the true function so as to obtain a signal-to-noise ratio of 1.0. The term "Mixture" in Table 1 indicates a mixture of Gaussian densities, $0.3\mathrm{N}(2, 1) + 0.7\mathrm{N}(-1, 0.5)$.

For the varying coefficient model (11), we set $\Phi_0 = \{\mathbf{1}\}$ to shrink f towards constant functions, whence the resulting model reduces to a linear regression model. Finally, we set $\Phi_0 = \{\mathbf{1}, Y, Y^2\}$ to shrink f towards the space of quadratic functions in (12), which results in the density p being shrunk towards the class of Gaussian distributions. We note that the prior for p in (12) is data-dependent.

Tables 1 shows the MSE of the posterior mean estimator for the varying coefficient model and the log-density model with sample sizes of $n = 200$ and 500. When the true function f belongs to the nominal parametric class, the posterior mean function resulting from the fHS prior outperforms the HS prior.

When the true function does not belong to the class of the parametric functions, the fHS prior performs comparably to the partial oracle estimator. The penalized spline method and the procedure based on the standard HS prior show smaller estimation error than the fHS prior and the partial oracle estimator (the standard B-spline estimator).

C Additional Real Data Examples of Linear Subspace Shrinkage for Additive Model

In this section, we examine the nonparametric additive model in low-dimensional settings when the component functions have linear forms. We apply the fHS prior to two well-known data sets: the first considers housing prices in Boston and the second concerns the progression of diabetes. The Boston housing data set has been previously analyzed in

Varying Coefficient Model												
Truth	Constant				Quadratic				Sine			
$n = 200$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$
Oracle	0.127(0.02)				1.474(0.05)	1.359(0.06)	1.895(0.08)	12.531(5.10)	1.094(0.05)	1.332(0.06)	1.893(0.08)	12.520(5.09)
HS	0.433(0.03)	0.640(0.04)	0.850(0.04)	2.437(0.07)	1.065(0.05)	1.258(0.06)	1.836(0.07)	4.832(0.11)	1.065(0.05)	1.383(0.06)	1.558(0.08)	5.370(0.15)
fHS1	0.171(0.02)	0.132(0.02)	0.129(0.02)	0.129(0.02)	1.471(0.05)	1.355(0.06)	1.876(0.08)	6.359(0.20)	1.113(0.05)	1.351(0.06)	1.895(0.08)	6.234(0.15)
fHS2	0.168(0.02)	0.131(0.02)	0.130(0.02)	0.128(0.02)	1.470(0.05)	1.354(0.06)	1.876(0.08)	6.360(0.20)	1.113(0.05)	1.350(0.06)	1.894(0.08)	6.235(0.15)
fHS3	0.167(0.02)	0.132(0.02)	0.129(0.02)	0.128(0.02)	1.471(0.05)	1.352(0.06)	1.876(0.08)	6.361(0.20)	1.113(0.05)	1.350(0.06)	1.895(0.08)	6.231(0.15)
$n = 500$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$
Oracle	0.058(0.01)				0.978(0.02)	0.508(0.03)	0.644(0.03)	2.231(0.06)	0.585(0.02)	0.484(0.03)	0.641(0.03)	2.230(0.06)
HS	0.189(0.02)	0.268(0.02)	0.345(0.03)	1.011(0.03)	0.601(0.02)	0.476(0.03)	0.668(0.03)	1.894(0.05)	0.481(0.02)	0.439(0.03)	0.602(0.03)	2.013(0.06)
fHS1	0.088(0.01)	0.061(0.01)	0.059(0.01)	0.058(0.01)	0.982(0.02)	0.513(0.03)	0.650(0.03)	2.205(0.06)	0.582(0.02)	0.478(0.03)	0.633(0.03)	2.158(0.06)
fHS2	0.087(0.01)	0.061(0.01)	0.059(0.01)	0.058(0.01)	0.982(0.02)	0.514(0.03)	0.649(0.03)	2.206(0.06)	0.583(0.02)	0.478(0.03)	0.633(0.03)	2.157(0.06)
fHS3	0.087(0.01)	0.061(0.01)	0.059(0.01)	0.058(0.01)	0.982(0.02)	0.514(0.03)	0.650(0.03)	2.205(0.06)	0.582(0.02)	0.478(0.03)	0.632(0.03)	2.159(0.06)
Log-density Model												
Truth	Normal				Log-normal				Mixture			
$n = 200$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 35$
Oracle	0.803(0.11)				5.227(0.46)	6.035(0.47)	9.949(0.65)	29.454(1.43)	11.920(0.60)	10.660(0.53)	11.888(0.87)	37.958(2.10)
HS	1.879(0.19)	2.335(0.20)	3.020(0.21)	9.187(0.42)	14.175(0.62)	6.725(0.45)	8.856(0.46)	17.860(0.64)	4.010(0.24)	3.737(0.21)	4.796(0.25)	12.254(0.37)
fHS1	2.118(0.20)	1.815(0.18)	1.780(0.17)	3.663(0.36)	12.234(0.54)	5.971(0.45)	5.348(0.47)	14.458(1.16)	4.705(0.31)	4.587(0.23)	5.326(0.26)	15.340(0.60)
fHS2	2.328(0.23)	1.844(0.18)	1.847(0.18)	2.048(0.30)	11.399(0.49)	5.231(0.39)	5.802(0.49)	11.990(0.90)	4.608(0.42)	4.085(0.23)	5.397(0.25)	14.497(0.35)
fHS3	2.292(0.21)	1.906(0.19)	1.852(0.17)	2.406(0.34)	10.344(0.51)	5.675(0.46)	5.356(0.41)	10.884(0.82)	4.321(0.28)	4.278(0.21)	5.286(0.24)	14.543(0.55)
$n = 500$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 5$	$k_n = 8$	$k_n = 11$	$k_n = 5$	$k_n = 8$	$k_n = 11$			
Oracle	0.328(0.03)				3.199(0.17)	3.416(0.18)	4.140(0.26)	10.504(1.43)	4.279(0.15)	5.129(0.17)	4.549(0.20)	8.882(0.27)
HS	0.743(0.06)	1.031(0.08)	1.299(0.08)	3.945(0.14)	7.029(0.28)	4.177(0.18)	4.074(0.20)	10.146(0.64)	2.564(0.12)	1.827(0.10)	2.299(0.11)	6.577(0.34)
fHS1	0.946(0.08)	0.948(0.08)	1.008(0.09)	1.317(0.12)	7.567(0.20)	3.966(0.19)	3.075(0.20)	14.423(1.16)	2.502(0.11)	2.273(0.13)	2.540(0.11)	7.303(0.84)
fHS2	0.940(0.08)	0.900(0.08)	0.916(0.07)	1.589(0.09)	8.125(0.21)	3.686(0.16)	3.101(0.20)	14.311(0.91)	2.665(0.14)	2.428(0.12)	2.591(0.11)	7.860(0.82)
fHS3	1.003(0.09)	0.950(0.07)	1.052(0.08)	1.610(0.12)	7.751(0.250	3.679(0.15)	3.148(0.18)	13.408(0.82)	2.647(0.13)	2.097(0.11)	2.629(0.11)	7.795(0.68)

Table 1: The results for varying coefficient models and log-density models. The description of this table is the same as Table 1 in the main article.

various places, including Buja et al. (1989), Breiman (1995), Lin & Zhang (2006) and Xue (2009). The data set is available in the `R` package `MASS`. The diabetes data set is famously used as a motivating example of Least Angle Regression (LARS; Efron et al. 2004) and it is contained in the `R` package `lars`.

The Boston housing data set contains the median value of 506 owner-occupied homes in the Boston area, together with several variables that might be associated with the median value. Using the standard notation for the variables in this data set, we assume a model of the following form:

$$\begin{aligned}\texttt{medv} &= \beta_0 + f_1(\texttt{crim}) + f_2(\texttt{indus}) + f_3(\texttt{nox}) + f_4(\texttt{rm}) + f_5(\texttt{age}) + f_6(\texttt{dis}) \\ &\quad + f_7(\texttt{tax}) + f_8(\texttt{ptratio}) + f_9(\texttt{b}) + f_{10}(\texttt{lstat}) + \epsilon,\end{aligned}$$

where $\epsilon \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_n)$. We also modeled the diabetes data set using each of the procedures that were applied to the housing data set. The diabetes data set consists of 19 variables measured on 403 patients, but we only considered continuous covariates and ignored missing samples by following the data pre-processing step suggested in Huang et al. (2012). The resulting data set contained 9 continuous variables and 366 samples. The response variable is Glycosolated hemoglobin (G.hem). The model applied to these data can be expressed as follows:

$$\begin{aligned}\texttt{G.hem} &= \beta_0 + f_1(\texttt{age}) + f_2(\texttt{bmi}) + f_3(\texttt{map}) + f_4(\texttt{tc}) + f_5(\texttt{ldl}) \\ &\quad + f_6(\texttt{tch}) + f_7(\texttt{ltg}) + f_8(\texttt{glu}) + f_9(\texttt{hdl}) + \epsilon.\end{aligned}$$

To compare the procedure based on the fHS priors to existing shrinkage and penalized likelihood methods, we considered the partially linear model that decomposes each component function f_j as a sum of linear and nonlinear parts as $f_j(X_j) = X_j\beta_{1j} + \Phi_j\beta_{2j}$, where $\beta_{1j} \in \mathbb{R}$ and $\beta_{2j} \in \mathbb{R}^{k_n}$ for $j = 1, \ldots, p$.

Huang et al. (2012) proposed a partially linear model with a group sparse penalty on β_{2j}'s. They considered the Minimax Concave Penalty (MCP; Zhang 2010) that can be expressed as $\lambda_1 \int_0^{\|\beta_{2j}\|_{n,2}} \{1 - x/(\lambda_0\lambda_1)\}_+ dx$, where λ_0 is the tuning parameter that controls the concavity of the penalty and λ_1 is the penalty parameter. We note that the group LASSO (Yuan & Lin 2006) is a special case of this class of penalties when $\lambda_0 = \infty$. The group LASSO was also included in our comparisons. For the real data analysis, we set $\lambda_0 = 1.1$ (Huang et al. 2012) and choose λ_1 by AIC and BIC. We used the HS prior in (16) on the spline coefficient β_{2j} for $j = 1, \ldots, p$. For the fHS priors, we imposed the prior in (14) for the additive model in (13). We defined Q_{0j} as the projection matrix of $\{\mathbf{1}, X_j\}$ for $j = 1, \ldots, p$.

For both data sets, each component function was modeled by B-spline bases, and 50 test data points were randomly selected to estimate the out-of-sample prediction error. For Bayesian methods, we generated $30,000$ posterior samples using the MCMC algorithm described in Section E in the supplementary material. Only the last $20,000$ samples were used in the analysis. We used multiple hyperparameters for the fHS prior, setting $b = \exp(-k_n \log n/10)$, $\exp(-k_n \log n/4)$, and $\exp(-k_n \log n/2)$, as in the previous simulation studies. For the penalized likelihood methods, AIC and BIC were used to choose tuning parameters. We also considered the least square estimator of the linear model and the standard B-spline estimator. Two hundred independent simulations of each procedure were used to generate Table 2 and Table 3. In these tables, "MSPE" refers to the average of the out-of-sample prediction errors and "LS" indicates the average number of linear components.

Table 2 summarizes the result for the Boston housing data set. For the considered hyperparameters, the procedure based on the fHS prior shows comparable or better performance than the other procedures. We note that the HS prior achieves a smaller prediction error than the fHS prior when $k_n = 11$. However, the variation of the prediction errors from different choice of k_n is much larger than for the fHS procedure. Also, the use of the HS prior does not provide model selection procedure, while the fHS prior contains a natural measure of model selection, as discussed in Section 3.2.

For the diabetes data, Table 3 shows that for all considered hyperparameters, the procedure based on the fHS prior attained smallest prediction errors. In this specific example, we note that AIC and BIC penalties were too strong, which led to all penalized likelihood procedures selecting only linear component functions.

	$k_n = 5$		$k_n = 8$		$k_n = 11$		$k_n = 35$	
	MSPE	LS	MSPE	LS	MSPE	LS	MSPE	LS
Linear	24.995(1.00)							
B-spline	14.250(0.70)		24.185(1.20)		177.673(22.93)		632.141(171.25)	
HS	16.166(0.81)		13.999(0.59)		**13.956**(0.56)		13.780(2.32)	
fHS1	13.934(0.62)	1.57	**13.940**(0.73)	2.85	14.425(0.98)	2.34	**13.609**(0.73)	4.74
fHS2	13.898(0.62)	1.67	13.980(0.72)	2.83	14.101(0.78)	2.13	14.804(1.01)	4.92
fHS3	**13.832**(0.59)	1.72	14.229(0.95)	2.91	13.989(0.76)	2.13	13.780(0.76)	4.90
GMCP (AIC)	14.000(0.58)	4.23	16.523(2.00)	3.12	15.674(1.27)	2.07	94.407(23.35)	2.50
GMCP (BIC)	15.288(0.685)	5.88	17.497(1.37)	6.81	19.536(1.79)	8.10	24.069(1.00)	10.00
GL (AIC)	14.287(0.65)	3.81	14.403(0.70)	3.94	14.523(0.69)	3.25	16.714(0.92)	5.40
GL (BIC)	15.220(0.75)	6.09	16.515(0.81)	6.15	17.130(0.81)	6.81	24.069(1.00)	10.00

Table 2: Boston data set. fHS1, fHS2, and fHS3 are the procedures based on the fHS prior with $b = \exp(-k_n \log n/10)$, $\exp(-k_n \log n/4)$, and $\exp(-k_n \log n/2)$, respectively. GMCP and GL are the penalized likelihood procedure based on the group MCP penalty and the group LASSO penalty, respectively. The smallest MSPE is noted in bold for each k_n.

D Proofs of Theorems

We first provide some lemmas that will be used to prove the main results.

Lemma D.1. *For arbitrary positive sequences u_n and w_n,*

$$\left(1 - \frac{u_n}{u_n + w_n}\right)^{u_n + w_n} \geq \exp\left\{-u_n + \frac{u_n^2}{2(u_n + w_n)}\right\}. \tag{1}$$

Proof. By Talyor's theorem, there exists $q_n^* \in (0, u_n/(u_n + w_n))$ such that

$$\begin{aligned}
\left(1 - \frac{u_n}{u_n + w_n}\right)^{u_n + w_n} &= \exp\left\{(u_n + w_n)\log\left(1 - \frac{u_n}{u_n + w_n}\right)\right\} \\
&= \exp\left\{(u_n + w_n)\left(-\frac{u_n}{u_n + w_n} + \frac{1}{(1 - q_n^*)^2}\frac{u_n^2}{2(u_n + w_n)^2}\right)\right\} \\
&\geq \exp\left\{-u_n + \frac{u_n^2}{2(u_n + w_n)}\right\}.
\end{aligned}$$

$\square$

	$k_n = 5$		$k_n = 8$		$k_n = 11$		$k_n = 35$	
	MSPE	LS	MSPE	LS	MSPE	LS	MSPE	LS
Linear	3147.478(54.9)							
B-spline	3445.47(66.9)		3750.42(91.6)		5432.97(758.3)		21777.20(1176.8)	
HS	3154.97(55.7)		3132.83(56.3)		3154.06(56.3)		4399.06(161.1)	
fHS1	3149.68(56.5)	5.37	3125.68(55.6)	7.11	**3054.99**(54.1)	7.17	4418.07(207.2)	8.98
fHS2	**3145.99**(56.3)	5.64	**3119.21**(55.9)	7.09	3065.60(54.9)	7.18	4388.90(161.5)	8.98
fHS3	3146.30(56.3)	5.57	3120.05(55.4)	7.08	3070.96(57.7)	7.17	4399.06(161.5)	8.98
GMCP (AIC)	3147.48(54.9)	9.00	3147.48(54.9)	9.00	3147.48(54.9)	9.00	**3147.48**(54.9)	9.00
GMCP (BIC)	3147.48(54.9)	9.00	3147.48(54.9)	9.00	3147.48(54.9)	9.00	**3147.48**(54.9)	9.00
GL (AIC)	3147.48(54.9)	9.00	3147.48(54.9)	9.00	3147.48(54.9)	9.00	**3147.48**(54.9)	9.00
GL (BIC)	3147.48(54.9)	9.00	3147.48(54.9)	9.00	3147.48(54.9)	9.00	**3147.48**(54.9)	9.00

Table 3: Diabetes data set. The description of this table is the same as Table 3.

Lemma D.2. *Suppose W follows a non-central chi-square distribution with m_n degrees of freedom and non-centrality parameter $\lambda_n \geq 0$, i.e, $W \sim \chi^2_{m_n}(\lambda_n)$. Also, let $w_n \to 0$ and $t_n \to \infty$ as $n \to \infty$ and assume that $m_n \prec t_n$. Then,*

$$P(W \leq \lambda_n w_n) \leq c_1 \lambda_n^{-1} \exp\{-\lambda_n(1-w_n)^2/8\}, \tag{2}$$

and

$$P(W > \lambda_n + t_n) \leq c_2 \left(\frac{t_n}{2m_n}\right)^{m_n/2} \exp\{m_n/2 - t_n/2\} + c_3 \lambda_n^{1/2} t_n^{-1} \exp\left\{-\frac{t_n^2}{32\lambda_n}\right\}, \tag{3}$$

where c_1, c_2, and c_3 are some positive constants.

Proof. W can be expressed as $W = \sum_{i=1}^{m_n}\{Z_i + (\lambda_n/m_n)^{1/2}\}^2$, where $Z_i \overset{i.i.d}{\sim} N(0,1)$ for $i = 1, \ldots, m$. Then, by the fact that $P(Z > a) \leq (2\pi)^{-1/2} a^{-1} \exp\{-a^2/2\}$ for any $a > 0$, we can show that there exist some positive constants c_1 such that

$$\begin{aligned} P\left(W \le \lambda_n w_n\right) &= P\Big\{\sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2}\sum_{i=1}^{m_n} Z_i + \lambda_n \le \lambda_n w_n\Big\} \\ &\le P\Big\{m_n^{-1/2}\sum_{i=1}^{m_n} Z_i \le -\lambda_n^{1/2}(1-w_n)/2\Big\} \\ &= P\Big\{|Z_1| \ge \lambda_n^{1/2}(1-w_n)/2\Big\}/2 \\ &\le c_1\lambda_n^{-1}\exp\{-\lambda_n(1-w_n)^2/8\}, \end{aligned}$$

since Z_1 follows a standard normal distribution.

By using Chernoffs's bound and the fact that $P(Z > a) \le (2\pi)^{-1/2}a^{-1}\exp\{-a^2/2\}$ for any $a > 0$, one can show that

$$\begin{aligned} P(W > \lambda_n + t_n) = P\left\{\sum_{i=1}^{m_n} Z_i^2 + 2(\lambda_n/m_n)^{1/2}\sum_{i=1}^{m_n} Z_i > t_n\right\} \\ \le\ P\left(\sum_{i=1}^{m_n} Z_i^2 > t_n/2\right) + P\left\{m_n^{-1/2}\sum_{i=1}^{m_n} Z_i > \lambda_n^{-1/2}t_n/4\right\} \\ \le\ c_2\left(\frac{t_n}{2m_n}\right)^{m_n/2}\exp\left\{m_n/2 - t_n/2\right\} + c_3\lambda_n^{1/2}t_n^{-1}\exp\left\{-\frac{t_n^2}{32\lambda_n}\right\}, \end{aligned}$$

where c_2 and c_3 are some positive constants. $\square$

Lemma D.3.

$$n\big\|Q_0\Phi\beta - Q_0Y\big\|_{n,2}^2/\sigma^2 \mid Y, \omega \sim \chi^2_{d_0},$$

and

$$n\big\|Q_1\Phi\beta - (1-\omega)Q_1Y\big\|_{n,2}^2/\{(1-\omega)\sigma^2\} \mid Y, \omega \sim \chi^2_{k_n-d_0}.$$

Proof. Recall that

$$\beta \mid Y, \omega \sim \mathrm{N}(\widetilde{\beta}_\omega, \widetilde{\Sigma}_\omega),$$

where

$$\widetilde{\beta}_\omega = \left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}}Y, \quad \widetilde{\Sigma}_\omega = \sigma^2\left(\Phi^{\mathrm{T}}\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)\Phi\right)^{-1}.$$

As shown in the proof of Lemma 3.1, $\Phi\left(\Phi^T\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}} = (1-\omega)Q_\Phi + \omega Q_0$, so

$$
\begin{gathered}
\mathbb{E}\left[Q_0\Phi\beta \mid Y,\omega\right] = Q_0 Y \\
\operatorname{Var}\left[Q_0\Phi\beta \mid Y,\omega\right] = \sigma^2 Q_0,
\end{gathered}
$$

which shows that $n\left\|Q_0\Phi\beta - Q_0Y\right\|_{n,2}^2/\sigma^2 \mid Y,\omega \sim \chi^2_{d_0}$.

Similarly,

$$
\begin{gathered}
\mathbb{E}\left[Q_1\Phi\beta \mid Y,\omega\right] = (1-\omega)Q_1 Y \\
\operatorname{Var}\left[Q_1\Phi\beta \mid Y,\omega\right] = \sigma^2(1-\omega) Q_1,
\end{gathered}
$$

which proves that $n\left\|Q_1\Phi\beta - (1-\omega)Q_1Y\right\|_{n,2}^2/\{(1-\omega)\sigma^2\} \mid Y,\omega \sim \chi^2_{k_n-d_0}$.

□

Proof of Lemma 3.1. As discussed in the paragraphs following Lemma 3.1 when $\mathfrak{L}(\Phi_0) \subsetneq \mathfrak{L}(\Phi)$, we can generate a new basis $\widetilde{\Phi} = [\Phi_0, \Phi_1]$ such that $\Phi_0^{\mathrm{T}}\Phi_1 = \mathbf{0}$ and $\mathfrak{L}(\Phi) = \mathfrak{L}(\widetilde{\Phi})$, which implies $Q_{\widetilde{\Phi}} = Q_\Phi$. Then,

$$
\begin{aligned}
& \Phi\left(\Phi^T\Phi + \frac{\omega}{1-\omega}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\right)^{-1}\Phi^{\mathrm{T}} \\
= \ & \widetilde{\Phi}\left(\widetilde{\Phi}^T\widetilde{\Phi} + \frac{\omega}{1-\omega}\widetilde{\Phi}^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\widetilde{\Phi}\right)^{-1}\widetilde{\Phi}^{\mathrm{T}} \\
= \ & [\Phi_0, \Phi_1]\begin{bmatrix} (\Phi_0^{\mathrm{T}}\Phi_0)^{-1} & \mathbf{0} \\ \mathbf{0} & (1-\omega)(\Phi_1^{\mathrm{T}}\Phi_1)^{-1} \end{bmatrix}\begin{bmatrix} \Phi_0^{\mathrm{T}} \\ \Phi_1^{\mathrm{T}} \end{bmatrix} \\
= \ & (1-\omega)Q_{\widetilde{\Phi}} + \omega Q_0 \\
= \ & (1-\omega)Q_\Phi + \omega Q_0.
\end{aligned}
$$

□

Proof of Lemma 3.2. From Polson & Scott (2012) it follows that

$$
\int_0^1 \omega^{A_n-1}(1-\omega)^{B_n-1}\exp\{-H_n\omega\}d\omega = \frac{\Gamma(A_n)\Gamma(B_n)}{\Gamma(A_n+B_n)}\exp\{-H_n\}\sum_{m=0}^{\infty}\frac{(A_n)_{(m)}}{(A_n+B_n)_{(m)}}\frac{{H_n}^m}{m!},
$$

where $(a)_{(m)} = a(a+1)\dots(a+m-1)$. We shall show that $\sum_{m=0}^{\infty}\left\{\frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}}\frac{{H_n}^m}{m!}\right\} \geq 1+Q_n^L$. By using Lemma D.1 and Stirling's approximation, i.e., $m! \asymp m^{m+1/2}\exp\{-m\}$, it follows that

$$
\begin{aligned}
& \sum_{m=0}^{\infty}\left\{\frac{(B_n)_{(m)}}{(A_n+B_n)_{(m)}}\frac{{H_n}^m}{m!}\right\} \\
= \;& 1+\frac{B_n}{A_n+B_n}\left\{H_n+\sum_{m=1}^{\infty}\left[\frac{(B_n+1)_{(m)}}{(A_n+B_n+1)_{(m)}}\frac{{H_n}^{m+1}}{(m+1)!}\right]\right\} \\
\geq \;& 1+\frac{B_n}{A_n+B_n}\left\{H_n+\sum_{m=1}^{\infty}\left[\frac{(B_n+m)!}{(A_n+B_n+m)!}\frac{{H_n}^{m+1}}{(m+1)!}\right]\right\} \\
\geq \;& 1+\frac{B_n}{A_n+B_n}\left\{H_n+D\sum_{m=1}^{\infty}\left[\left(\frac{B_n+m}{A_n+B_n+m}\right)^{A_n+B_n+m+1/2}(B_n+m)^{-A_n}\right.\right. \\
& \left.\left. e^{A_n}\frac{{H_n}^{m+1}}{(m+1)!}\right]\right\} \\
\geq \;& 1+\frac{B_n}{A_n+B_n}\left\{H_n+D\sum_{m=1}^{T_n}\left[\left(\frac{B_n+1}{A_n+B_n+1}\right)^{1/2}(B_n+m)^{-A_n}\right.\right. \\
& \left.\left.\times\left(\frac{B_n+m}{A_n+B_n+m}\right)^{A_n+B_n+m} e^{A_n}\frac{{H_n}^{m+1}}{(m+1)!}\right]\right\} \\
\geq \;& 1+\frac{B_n}{A_n+B_n}\left\{H_n+D\left(\frac{B_n+1}{A_n+B_n+1}\right)^{1/2}(B_n+T_n)^{-A_n}\right. \\
& \left.\times\exp\left\{\frac{A_n^2}{2(A_n+B_n+T_n)}\right\}\sum_{m=2}^{T_n+1}\frac{{H_n}^m}{m!}\right\},
\end{aligned}
\tag{4}
$$

where $T_n = \max\{A_n^2, 3\lceil H_n\rceil\}$, and D is some positive constant.

Since $H_n < (T_n+2)\exp\{1\}$, by using the Stirling's approximation, the term $\sum_{m=2}^{T_n+1} H_n/m!$

in (4) can be expressed as follows:

$$
\begin{aligned}
\sum_{m=2}^{T_n+1} \frac{H_n^m}{m!} &= \exp\{H_n\} - 1 - H_n - \sum_{m=T_n+2}^{\infty} \frac{H_n^m}{m!} \\
&\preceq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \sum_{m=T_n+2}^{\infty} \left(\frac{\exp\{1\}H_n}{T_n+2} \right)^m \\
&\leq \exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2}
\end{aligned}
$$

Therefore, (4) can be bounded by

$$
\begin{aligned}
& 1 + \frac{B_n}{A_n+B_n} \Bigg\{ H_n + D \left(\frac{B_n+1}{A_n+B_n+1} \right)^{1/2} (B_n+T_n)^{-A_n} \\
& \times \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+ \Bigg\} \\
\geq \;& 1 + \frac{B_n H_n}{A_n+B_n} + \frac{DB_n}{(A_n+B_n)^{3/2}} (B_n+T_n)^{-A_n} \left(\exp\{H_n\} - 1 - H_n - (T_n+2)^{-1/2} \right)_+ ,
\end{aligned}
$$

where $(\cdot)_+$ denotes the positive hinge function (i.e., for any $t \in \mathbb{R}$, $(t)_+ = t$, if $t > 0$, and $(t)_+ = 0$, otherwise).

Also, since $(B_n+m)!/(A_n+B_n+m)! < 1$ for any positive integer m, it follows that

$$
H_n + \sum_{m=1}^{\infty} \left[\frac{(B_n+m)!}{(A_n+B_n+m)!} \frac{{H_n}^{m+1}}{(m+1)!} \right] \leq \exp\{H_n\},
$$

which completes the proof. □

Proof of Theorem 3.3. Let β^* denote the projection of the true F_0 on the basis $\{\phi_j\}_{1\leq j\leq k_n}$, i.e.,

$$
\beta^* = \operatorname{argmin}_{\beta\in\mathbb{R}^{k_n}} \left\| F_0 - \Phi\beta \right\|_{2,n}. \tag{5}
$$

We shall treat β^* as the *pseudo-true* parameter and study the posterior concentration of $\Phi\beta$ in the posterior around $\Phi\beta^*$.

To prove Theorem 3.3, it is sufficient to show that the posterior probability in the equation $\mathbb{E}_0\left[P\left\{\left\|\Phi\beta - F_0\right\|_{2,n} > M_n(f_0)^{1/2} \mid Y\right\}\right]$ converges in probability to zero. The posterior probability in the expectation can be decomposed as follows:

$$\begin{aligned}
& P\left[\left\|\Phi\beta - F_0\right\|_{n,2} > M_n^{1/2} \mid Y\right] \\
\leq\ & P\left[\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y\right] + \mathbb{1}\left[\left\|\Phi\beta^* - F_0\right\|_{n,2} > M_n^{1/2}/2\right],
\end{aligned}$$

where β^* is defined in (5) and $\mathbb{1}(\cdot)$ is the indicator function. The second term on the right-hand side of this expression is always zero when $F_0 \in \mathfrak{L}(\Phi_0)$, since we assume that the column space of Φ_0 is contained in the column space of Φ, and its expectation with respect to the true density is asymptotically zero when $F_0^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F_0 \asymp n$ from (9). Therefore, we focus on the first term on the right-hand side. Since $\Phi\beta = Q_1\Phi\beta + Q_0\Phi\beta$, by Lemma 3.1. the first term can be decomposed as

$$\begin{aligned}
& P\left[\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y\right] = E_{\omega|Y}\left[P\left(\left\|\Phi\beta - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/2 \mid Y, \omega\right)\right] \\
\leq\ & \mathbb{E}_{\omega|Y}\left[P\left(\left\|\Phi\beta - \Phi\widetilde{\beta}_\omega\right\|_{n,2} > M_n^{1/2}/4 \mid Y, \omega\right)\right] \\
& + \mathbb{E}_{\omega|Y}\left[P\left(\left\|\Phi\widetilde{\beta}_\omega - \Phi\beta^*\right\|_{n,2} > M_n^{1/2}/4 \mid Y, \omega\right)\right] \\
\leq\ & \mathbb{E}_{\omega|Y}\left[P\left(\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right)\right] \\
& + \mathbb{E}_{\omega|Y}\left[P\left(\left\|Q_1\Phi\beta^* - (1-\omega)Q_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right)\right] \\
& + \mathbb{E}_{\omega|Y}\left[P\left(\left\|Q_0\Phi\beta - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right)\right] \\
& + \mathbb{1}\left[\left\|Q_0\Phi\beta^* - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8\right],
\end{aligned}$$

where $\Phi\widetilde{\beta}_\omega = (1-\omega)\mathrm{Q}_\Phi Y + \omega\mathrm{Q}_0 Y = (1-\omega)\mathrm{Q}_1 Y + \mathrm{Q}_0 Y$.

We denote

$$\begin{aligned}
W_1 &= P\left(\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right), \\
W_2 &= P\left(\left\|Q_1\Phi\beta^* - (1-\omega)Q_1 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right), \\
W_3 &= P\left(\left\|Q_0\Phi\beta - Q_0 Y\right\|_{n,2} > M_n^{1/2}/8 \mid Y, \omega\right).
\end{aligned}$$

The indicator function in the fourth term converges to zero in probability, since $\left\|Q_0 Y - Q_0\Phi\beta^*\right\|_{2,n}^2$ achieves the parametric optimal rate. To complete the proof we show that the expectations of W_1, W_2, and W_3 with respect to the marginal posterior distribution of ω converge to zero in probability.

First consider W_3. Since $n\left\|\mathrm{Q}_0\Phi\beta - \mathrm{Q}_0 Y\right\|_{2,n}^2/\sigma^2 \mid Y,\omega \sim \chi^2_{d_0}$ by Lemma D.3, by using Lemma D.2 it follows that

$$\begin{aligned} E_{\omega|Y}\left[W_3\right] &= E_{\omega|Y}\left[P\left\{\left\|\mathrm{Q}_0\Phi\beta - \mathrm{Q}_0 Y\right\|_{2,n} > M_n^{1/2}/8 \mid Y,\omega\right\}\right] \\ \leq\ & C\left(\frac{nM_n}{64\sigma d_0}\right)^{d_0/2}\exp\{-nM_n/(128\sigma^2)\}, \end{aligned}$$

for some constant C.

The last quantity converges to zero as n tends to ∞, which implies that $\mathbb{E}_{\omega|Y}[W_3] = o_p(1)$. Now we obtain the bounds on W_1. By Lemma D.3 $n\left\|\mathrm{Q}_1\Phi\beta - (1-\omega)\mathrm{Q}_1 Y\right\|_{2,n}^2/\{(1-\omega)\sigma^2\} \mid Y \sim \chi^2_{k_n-d_0}$. By using Lemma D.2, it follows that

$$\begin{aligned} W_1 \ \leq\ & \left[\frac{nM_n}{64\sigma^2(k_n-d_0)}(1-\omega)^{-1}\right]^{\frac{k_n-d_0}{2}}\exp\left\{\frac{k_n-d_0}{2} - \frac{nM_n}{128\sigma^2}(1-\omega)^{-1}\right\} \\ & \times\mathbb{1}\left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} > k_n - d_0\right] + \mathbb{1}\left[\frac{nM_n}{64\sigma^2}(1-\omega)^{-1} \leq k_n - d_0\right]. \end{aligned}$$

We denote the two terms in this expression as $W_{1,1}$ and $W_{1,2}$.

By using Lemma 3.2 and defining $\widehat{\omega} = (k_n - d_0)/\{nM_n/(64\sigma^2) + k_n - d_0\}$, it follows that

$$
\begin{aligned}
&\mathbb{E}_{\omega|Y}\left[W_{1,1}\right] \\
= \ & \frac{1}{m(Y)}\left[\frac{nM_n\exp\{1\}}{64\sigma^2(k_n-d_0)}\right]^{\frac{k_n-d_0}{2}}\int_{m_n}^1 \omega^{a+\frac{k_n-d_0}{2}-1}(1-\omega)^{b-\frac{k_n-d_0}{2}-1} \\
& \times\exp\left\{-\frac{nM_n}{128\sigma^2}(1-\omega)^{-1}-H_n\omega\right\}d\omega \\
\leq \ & \frac{1}{m(Y)}\left[\frac{nM_n\exp\{1\}}{64\sigma^2(k_n-d_0)}\right]^{\frac{k_n-d_0}{2}}\int_{m_n}^1 \omega^{a-1}(1-\omega)^{b-1}\exp\left\{-H_n\omega\right\}d\omega \\
& \times\widehat{\omega}^{\frac{k_n-d_0}{2}}(1-\widehat{\omega})^{-\frac{k_n-d_0}{2}}\exp\left\{-\frac{nM_n}{128\sigma^2}(1-\widehat{\omega})^{-1}\right\} \\
= \ & \frac{1}{m(Y)}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\}\int_{m_n}^1 \omega^{a-1}(1-\omega)^{b-1}\exp\left\{-H_n\omega\right\}d\omega, \qquad (6)
\end{aligned}
$$

where $m_n=\max[0,1-nM_n/\{16\sigma^2(k_n-d_0)\}]$.

Also,

$$
\begin{aligned}
&\mathbb{E}_{\omega|Y}[W_{1,2}] = P_{\omega|Y}\left[\omega<1-\frac{nM_n}{64\sigma^2(k_n-d_0)}\right] \\
= \ & \frac{1}{m(Y)}\int_0^{1-\frac{nM_n}{64\sigma^2(k_n-d_0)}}\omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\
\leq \ & \frac{1}{m(Y)}\left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1}\int_0^1 \omega^{a+(k_n-d_0)/2-1}\exp\{-H_n\omega\}d\omega \\
\leq \ & \left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1}\frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(a+(k_n-d_0)/2)\Gamma(b)}H_n^{-1}\mathbb{1}\left(1-\frac{nM_n}{64\sigma^2(k_n-d_0)}\geq 0\right) \\
& \times\exp\{H_n\}\left[1+\frac{bH_n}{a+b+(k_n-d_0)/2}+D\frac{b(b+T_n)^{-a-(k_n-d_0)/2}}{(a+b+(k_n-d_0)/2)^{3/2}}\right. \\
& \left.\times\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right]^{-1}, \qquad (7)
\end{aligned}
$$

where $T_n=\max\{(a+(k_n-d_0)/2)^2,3\lceil H_n\rceil\}$ and D is some constant.

We now consider two cases: (i) when $F_0 \in \mathfrak{L}(\Phi_0)$ and (ii) when $F_0^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F_0 \asymp n$.

Case (i) $F_0 \in \mathfrak{L}(\Phi_0)$: Recall that in this case $M_n = \zeta_n n^{-1}$ for any arbitrary diverging sequence ζ_n. First, we show that $\mathbb{E}_{\omega|Y}[W_1] \xrightarrow{p} 0$ by proving that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$ and $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{p} 0$.

Applying Lemma 3.2, it follows that (6) is bounded above by

$$\begin{aligned} \mathbb{E}_{\omega|Y}\left[W_{1,1}\right] &\leq \frac{C\exp\left\{-nM_n/(128\sigma^2)\right\}\left(1+\frac{b}{a+b}\exp\{H_n\}\right)}{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+} \\ \leq\ & C\exp\left\{-\frac{nM_n}{128\sigma^2}\right\}\left(1+\frac{b}{a+b}\exp\{H_n\}\right), \end{aligned} \tag{8}$$

where $\delta_n = bH_n/(a+b+(k_n-d_0)/2)$ and $u_n = (a+b)(b+T_n)^{-a_n-(k_n-d_0)/2}/(a+b+(k_n-d_0)/2)^{3/2}$ with $T_n = \max\{(a+(k_n-d_0)/2)^2, 3\lceil H_n\rceil\}$, and C and D are some constants.

Since $2H_n \sim \chi^2_{k_n-d_0}$, by Lemma D.2 and defining $q_n = k_n^{-1/2}(\log k_n)^{1/2}(-\log b)^{1/2}$, it follows that

$$P\left[H_n > k_n q_n/2\right] \leq \exp\{-ck_n q_n\}, \tag{9}$$

for some constant c. Hence, by the condition that $k_n \log k_n \prec -\log b$, it is clear that $b\exp\{H_n\} = o_p(1)$, which shows that $\mathbb{E}_{\omega|Y}[W_{1,1}] = o_p(1)$.

Similarly, since $\Gamma(b)^{-1} \asymp b$, (7) is bounded by

$$C' b\exp\{H_n\}\left(\frac{nM_n}{64\sigma^2(k_n-d_0)}\right)^{b-1},$$

for some constant C'. By (9), $b\exp\{H_n\} = o_p(1)$, which implies $\mathbb{E}_{\omega|Y}[W_{1,2}] = o_p(1)$.

We next show that $E_{\omega|Y}[W_2]$ converges in probability to zero. Applying Lemma 3.2, it follows that

$$\begin{aligned}
& \mathbb{E}_{\omega|Y}[W_2] = \mathbb{E}_{\omega|Y}\left[P\left[\left\|(1-\omega)Q_1Y - Q_1\Phi\beta^*\right\|_{n,2} > M_n^{1/2}/8 \mid Y,\omega\right]\right] \\
= \;& P_{\omega|Y}\left[\omega < 1 - \left(\frac{nM_n}{64\sigma^2 H_n}\right)^{1/2}\right] \\
= \;& \frac{1}{m(Y)}\int_0^{1-\left(\frac{nM_n}{128\sigma^2H_n}\right)^{1/2}} \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\
\leq \;& \mathbb{1}\left\{1 - \left(\frac{nM_n}{128\sigma^2H_n}\right)^{1/2} \geq 0\right\}\frac{1}{m(Y)}\left(\frac{nM_n}{64\sigma^2H_n}\right)^{(b-1)/2} \\
& \times \int_0^1 \omega^{a+(k_n-d_0)/2-1}\exp\{-H_n\omega\}d\omega \\
\leq \;& \mathbb{1}\left\{1 - \left(\frac{nM_n}{128\sigma^2H_n}\right)^{1/2} \geq 0\right\}\frac{\Gamma(a+b+(k_n-d_0)/2)}{\Gamma(b)\Gamma(a+(k_n-d_0)/2)}\left(\frac{nM_n}{128\sigma^2H_n}\right)^{(b-1)/2} \\
& \times \exp\{H_n\}\left\{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right\}^{-1} \\
\leq \;& Cb\left(\frac{nM_n}{128\sigma^2}\right)^{(b-1)/2} H_n^{1/2}\exp\{H_n\},
\end{aligned}$$

where C is some constant, and δ_n and u_n are defined following (8).

From (9), it follows that $b\{nM_n/(128\sigma^2)\}^{(b-1)/2}H_n^{1/2}\exp\{H_n\}$ is bounded by $b\{nM_n/(128\sigma^2)\}^{(b-1)/2}(k_nq_n/2)^{1/2}\exp\{k_nq_n/2\}$ with probability greater than $1-\exp\{-ck_nq_n\}$ from which it follows that $\mathbb{E}_{\omega|Y}[W_2] = o_p(1)$.

Case (ii) $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp n$:

Recall that in this case $M_n = \zeta_n n^{-2\alpha/(1+2\alpha)}\log n$ for any arbitrary diverging sequence ζ_n, and δ_n and u_n are defined following (8). From (6) it follows that

$$\begin{aligned}
& \mathbb{E}_{\omega|Y}\left[W_{1,1}\right] \leq \frac{1}{m(Y)}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\}\int_{m_n}^1 \omega^{a-1}(1-\omega)^{b-1}\exp\left\{-H_n\omega\right\}d\omega \\
\leq \;& C\exp\left\{-\frac{nM_n}{128\sigma^2}\right\}\frac{1+\frac{b}{a+b}\exp\{H_n\}}{1+\delta_n+u_n\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+},
\end{aligned}$$

for some constant C.

By Lemma D.2, for any sequence $w_n \to 0$, H_n is larger than $w_n F_0^{\mathrm{T}} Q_1 F_0/\sigma^2$ with probability greater than $1 - \exp\{-cF_0^{\mathrm{T}} Q_1 F_0(1-w_n)^2/\sigma^2\}$ for some constant c. Since $F_0^{\mathrm{T}}(\mathrm{I} - \mathrm{Q}_0)F_0 \asymp n$ implies $F_0^{\mathrm{T}} Q_1 F_0 \asymp n$, the last line in the above display can be expressed as

$$C' \exp\left\{-\frac{nM_n}{128\sigma^2}(k_n - d_0)^{3/2}(b + T_n)^{(k_n-d_0)/2}\right\} + o_p(1),$$

where $T_n = \max\{(a + (k_n - d_0)/2)^2, 3H_n\}$ and C' is some positive constant. Therefore, to show $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$, it is sufficient to prove that $T_n^{(k_n-d_0)/2}\exp\{-nM_n/(128\sigma^2)\} = o_p(1)$. For any $\epsilon > 0$,

$$\begin{aligned} & P\left[T_n^{(k_n-d_0)/2}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\} > \epsilon\right] \\ \leq\ & P\left[(3H_n)^{(k_n-d_0)/2}\exp\left\{-\frac{nM_n}{128\sigma^2}\right\} > \epsilon\right] + P\left[3H_n < (a + (k_n - d_0)/2)^2\right] \\ \leq\ & P\left[\log H_n > \zeta_n \log n\right] + P\left[3H_n < (a + (k_n - d_0)/2)^2\right]. \end{aligned}$$

Since $\zeta_n \to \infty$ as n tends to ∞, from (3) in Lemma D.2, it follows that the first term in the above display can be bounded above by $\exp\{-c'(n_n^{\zeta} - F_0^{\mathrm{T}} Q_1 F_0/\sigma^2)\}$ for some constant c'. Similarly, from (2) in Lemma D.2, the second term is bounded by $\exp\{-c'' F_0^{\mathrm{T}} Q_1 F_0/\sigma^2\}$ with some constant c'', which proves that $\mathbb{E}_{\omega|Y}[W_{1,1}] \xrightarrow{p} 0$.

Since $nM_n \succ k_n$, the indicator function $\mathbb{1}(1 - nM_n/(64\sigma^2(k_n - d_0)) \geq 0)$ in (7) is zero when n is large enough, which results in $\mathbb{E}_{\omega|Y}[W_{1,2}] \xrightarrow{p} 0$.

The marginal posterior mean of W_2 can be decomposed as

$$\begin{aligned} \mathbb{E}_{\omega|Y}[W_2] \ \leq\ & P_{\omega|Y}\left[\left\|(1-\omega)\mathrm{Q}_1 Y - \mathrm{Q}_1 Y\right\|_{n,2} > \frac{1}{16}M_n^{1/2}\right] \\ & + \mathbb{1}\left[\left\|\mathrm{Q}_1 Y - \mathrm{Q}_1 \Phi\beta^*\right\|_{n,2} > \frac{1}{16}M_n^{1/2}\right]. \end{aligned}$$

Results provided by Zhou et al. (1998) (see equation (9) on page 10) show that the second term in the previous expression is $o_p(1)$. The first term can be expressed as

$$\begin{aligned}
& P_{\omega|Y}\left[\omega > \left(\frac{nM_n}{256\sigma^2 H_n}\right)^{1/2}\right] \\
& = \frac{1}{m(Y)}\int_{\left(\frac{nM_n}{256\sigma^2 H_n}\right)^{1/2}}^{1} \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}\exp\{-H_n\omega\}d\omega \\
\leq \; & \frac{1}{m(Y)}\exp\left\{-H_n^{1/2}\left(nM_n/(256\sigma^2)\right)^{1/2}\right\}\int_0^1 \omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b-1}d\omega \\
\leq \; & \left[u_n\exp\{-H_n\}\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right]^{-1} \\
& \times\exp\left\{-H_n^{1/2}\left(nM_n/(256\sigma^2)\right)^{1/2}\right\},
\end{aligned}$$

for some positive constant D. Since $H_n/n = O_p(1)$ and $-\log b \prec n^{1/2}k_n^{1/2}$, the above quantity converges in probability to zero, which completes the proof. $\square$

Proof of Theorem 3.4. We shall prove the result by separating two cases that are $F_0 \in \mathfrak{L}(\Phi_0)$ and $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp 0$.

Case (i) $F_0 \in (\Phi_0)$: We use the formulation in (7). By plugging $M_n = 64\sigma^2(k_n-d_0)S_{0,n}/n$ in (7), it follows that

$$P\left(\omega < 1-S_{0,n} \mid Y\right) \leq Cb\exp\{H_n\}\left(b^{1-\epsilon_0}/k_n\right)^{b-1},$$

for some constant $C>0$. Since $k_n b^{\epsilon_0}\exp\{H_n\} = o_p(1)$ by (9), $P(\omega < 1-S_{0,n} \mid Y) = o_p(1)$.

Case (ii) $F_0^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)F_0 \asymp 0$: By following the formulation in (8), it follows that

$$\begin{aligned}
& P\left(\omega > S_{1,n} \mid Y\right) = \frac{1}{m(Y)}\int_{S_{1,n}}^{1}\omega^{a+(k_n-d_0)/2-1}(1-\omega)^{b}\exp\{-H_n\omega\}d\omega \\
\leq \; & \left[u_n\exp\{-H_n\}\frac{Db}{a+b}\left(\exp\{H_n\}-1-H_n-(T_n+2)^{-1/2}\right)_+\right]^{-1}\exp\left\{-H_nS_{1,n}\right\} \\
\leq \; & Cb^{-1}\exp\{-H_nS_{1,n}\},
\end{aligned}$$

for some constant $C>0$. Since $H_n/n = D_0 + o_p(1)$ for some constant $D_0>0$, $b^{-1}\exp\{-H_nS_{1,n}\} = o_p(1)$, which completes the proof. $\square$

E Computation Strategy: Slice Sampling

In model (1), the conditional posterior distribution of τ based on the fHS prior can be expressed as

$$\pi(\tau \mid Y, \beta) \propto (\tau^2)^{-(k_n-d_0)/2+b-1/2}(1+\tau^2)^{-a-b}\exp\{-\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta/(2\sigma^2)\}.$$

By reparameterizing $\eta = 1/\tau^2$, the resulting conditional posterior distribution of η can be expressed as

$$\pi(\eta \mid Y, \beta) \propto \eta^{a+(k_n-d_0)/2-1}\exp\{-\beta^{\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta/(2\sigma^2)\}\frac{1}{(1+\eta)^{a+b}}.$$

As in Polson et al. (2014), a slice sampling method (Neal 2003) can be used to sample η from its conditional posterior distribution. The resulting MCMC algorithm is described in Algorithm 1.

Algorithm 1 MCMC algorithm for simple nonparametric regression models

Choose an initial value $\beta^{(0)}$ and $\tau^{(0)}$.
For l in $0 : (L-1)$
Sample $\beta^{(l+1)}$ from $\mathrm{N}(\widetilde{\beta}_{\omega^{(l)}}, \sigma^2\widetilde{\Sigma}_{\omega^{(l)}})$, where $\widetilde{\beta}_{\omega}$ and $\widetilde{\Sigma}_{\omega}$ are defined in (7).
(Slice sampling step) Set $\eta = 1/\tau^{2(l)}$ and $t = (\eta+1)^{-a-b}$.
Sample $u \sim Unif(0,t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.
Sample $\eta^* \sim truncated\ Gamma(a+(k_n-d_0)/2, \beta^{(l+1)\mathrm{T}}\Phi^{\mathrm{T}}(\mathrm{I}-\mathrm{Q}_0)\Phi\beta^{(l+1)}/(2\sigma^2))$ on $(0,t^*)$,
Update $\tau^{(l+1)}$ by $\eta^{*-1/2}$.
End.

In the additive model in (13) with a product of the fHS priors, the conditional posterior distribution of β_j given ω_j and the other coefficients $\beta_{(-j)}$, for $j = 1, \ldots, p$, can be expressed as

$$\beta_j \mid \omega_j, \beta_{(-j)}, Y \sim \mathrm{N}\left(\widetilde{\beta}_{j,\omega}, \sigma^2\widetilde{\Sigma}_{j,\omega}\right),$$

where

$$\widetilde{\beta}_{j,\omega} = \widetilde{\Sigma}_{j,\omega}\Phi_j^{\mathrm{T}} r_j, \quad \widetilde{\Sigma}_{j,\omega} = (1-\omega_j)\left(\Phi_j^{\mathrm{T}}\Phi_j\right)^{-1}, \quad r_j = Y - \sum_{l\neq j}\Phi_l\beta_l. \tag{10}$$

It follows that sampling Algorithm 1 can be extended to additive regression models to obtain Algorithm 2 below.

Algorithm 2 MCMC algorithm for additive regression models

Choose an initial value $\beta_j^{(0)}$ and $\tau_j^{(0)}$ for $j = 1, \cdots, p$.
For l in $0 : (L-1)$
  For j in $1 : p$
    Sample $\beta_j^{(l+1)}$ from $\mathrm{N}(\widetilde{\beta}_{j,\omega^{(l)}}, \sigma^2\widetilde{\Sigma}_{j,\omega^{(l)}})$, where $\widetilde{\beta}_{j,\omega}$ and $\widetilde{\Sigma}_{j,\omega}$ are defined in (10).
  End.
  For j in $1 : p$
  (Slice sampling step)
    Set $\eta = 1/\tau_j^{2(l)}$ and $t = (\eta+1)^{-a-b}$.
    Sample $u \sim Unif(0,t)$ and set $t^* = u^{-(a+b)^{-1}} - 1$.
    Sample $\eta^* \sim truncated\ Gamma(a + k_n/2, \beta_j^{(l+1)\mathrm{T}}\Phi_j^{\mathrm{T}}\Phi_j\beta_j^{(l+1)}/(2\sigma^2))$ on $(0, t^*)$,
    Update $\tau_j^{(l+1)}$ by $\eta^{*-1/2}$.
  End.
End.

The computational complexity of Algorithm 2 for each iteration is $O(pk_n^3) + O(npk_n)$. The term $O(pk_n^3)$ arises from updating the p blocks of β, each of length k_n. The joint update of β without separating into blocks is also available, but it requires the inversion of a $pk_n \times pk_n$ matrix. Even though this joint update may improve the convergence of the MCMC chain, its computational burden for each iteration will significantly increase to $O(p^3k_n^3)$. While Bhattacharya et al. (2016) proposed a procedure reducing this complexity to $O(n^2pk_n)$ by avoiding the matrix inversion step, we stick to the block-wise update pro-

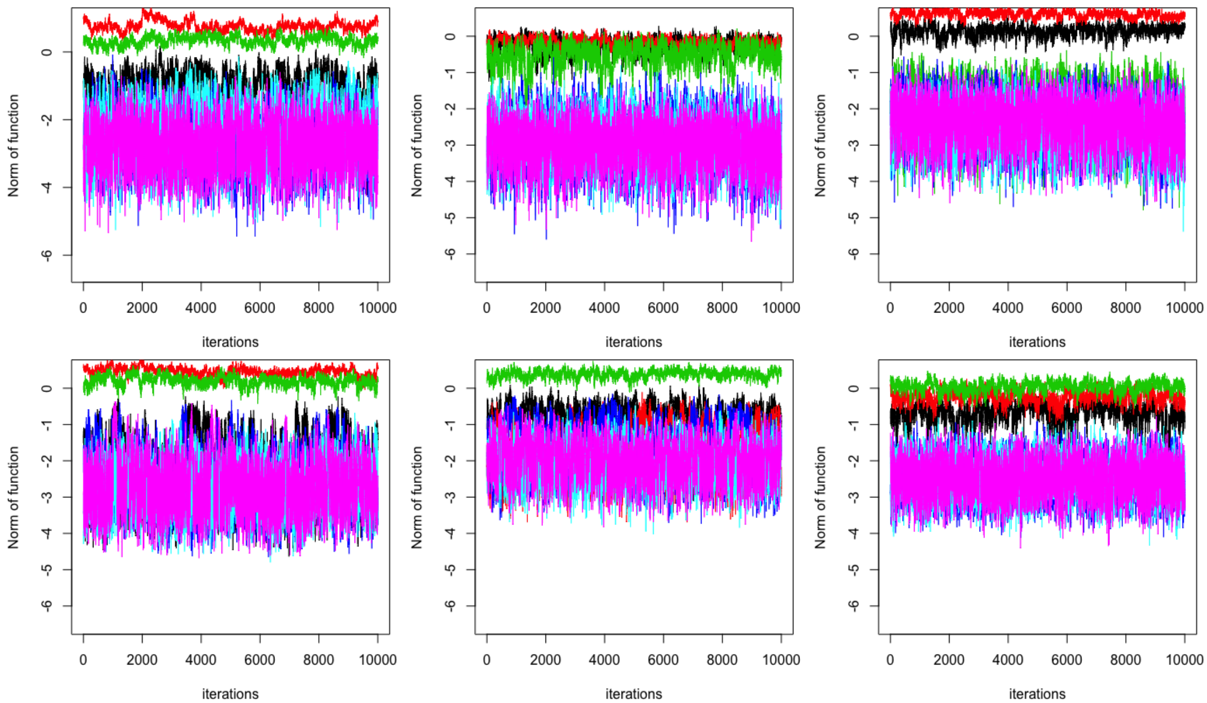

Figure 1: Trace plots. The first and the second row are cases when $k_n = 11$ and $k_n = 35$, respectively. Scenario 1, 2, and 3 are illustrated in the left, middle, and right column, respectively.

cedure in Algorithm 2, and its empirical performance was promising in various simulation and real data analysis.

F Trace Plots for Simulation Scenarios

In this section, we examine some trace plots of simulated data sets considered in Section 4 in the main article.

We examine the mixing behavior of the fHS procedure in the additive model context. We selected six component functions, three of which were null while the other three non-null. Each sub-plot of Figure 1 shows the trace plots of the empirical L_2 norms of these six functions, with the different functions color-coded. The different columns correspond to the three simulation scenarios, while the top and bottom rows correspond to $k_n = 11$ and 35 respectively. The mixing in all the cases seems reasonable from examination of the trace plots, and no obvious difference is potted between $k_n = 11$ and $k_n = 35$.

References

Bhattacharya, A., Chakraborty, A. & Mallick, B. K. (2016), 'Fast sampling with Gaussian scale-mixture priors in high-dimensional regression', *Biometrika* **103**(4), 985–991.

Breiman, L. (1995), 'Better subset regression using the nonnegative garrote', *Technometrics* **37**(4), 373–384.

Buja, A., Hastie, T. & Tibshirani, R. (1989), 'Linear smoothers and additive models', *Annals of Statistics* **17**(2), 453–510.

Claeskens, G., Krivobokova, T. & Opsomer, J. D. (2009), 'Asymptotic properties of penalized spline estimators', *Biometrika* **96**(3), 529–544.

De Boor, C. (2001), *A Practical Guide to Splines*, revised edn, Springer, Newyork, chapter 9.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Huang, J., Wei, F. & Ma, S. (2012), 'Semiparametric regression pursuit', *Statistica Sinica* **22**(4), 1403.

Lin, Y. & Zhang, H. H. (2006), 'Component selection and smoothing in multivariate nonparametric regression', *Annals of Statistics* **34**(5), 2272–2297.

Neal, R. M. (2003), 'Slice sampling', *Annals of Statistics* **31**(3), 705–767.

Polson, N. G. & Scott, J. G. (2012), 'On the half-Cauchy prior for a global scale parameter', *Bayesian Analysis* **7**(4), 887–902.

Polson, N. G., Scott, J. G. & Windle, J. (2014), 'The Bayesian bridge', *Royal Statistical Society: Series B* **76**(4), 713–733.

Xue, L. (2009), 'Consistent variable selection in additive models', *Statistica Sinica* **19**, 1281–1296.

Yuan, M. & Lin, Y. (2006), ‘Model selection and estimation in regression with grouped variables’, *Royal Statistical Society: Series B* **68**(1), 49–67.

Zhang, C.-H. (2010), ‘Nearly unbiased variable selection under minimax concave penalty’, *Annals of Statistics* **38**(2), 894–942.

Zhou, S., Shen, X. & Wolfe, D. (1998), ‘Local asymptotics for regression splines and confidence regions’, *Annals of Statistics* **26**(5), 1760–1782.