Journal of Machine Learning Research 26 (2025) 1-38 Submitted 3/22; Revised 6/25; Published 7/25

Bayesian Scalar-on-Image Regression with a Spatially Varying Single-layer Neural Network Prior

Ben Wu[*] WUBEN@RUC.EDU.CN
Center for Applied Statistics, School of Statistics
Renmin University of China
Beijing, China

Keru Wu[*] KERU.WU@DUKE.EDU
Department of Statistical Sciences
Duke University
Durham, NC 27708, USA

Jian Kang[†] JIANKANG@UMICH.EDU
Department of Biostatistics
University of Michigan
Ann Arbor, MI 48109, USA

Editor: Debdeep Pati

Abstract

Deep neural networks (DNN) have been widely used in scalar-on-image regression to predict an outcome variable from imaging predictors. However, training DNN typically requires large sample sizes for accurate prediction, and the resulting models often lack interpretability. In this work, we propose a novel Bayesian nonlinear scalar-on-image regression framework with a spatially varying single-layer neural network (SV-NN) prior. The SV-NN is constructed using a single hidden layer neural network with its weights generated by the soft-thresholded Gaussian process. Our framework enables the selection of interpretable image regions while achieving high prediction accuracy with limited training samples. The SV-NN offers large prior support for the imaging effect function, facilitating efficient posterior inference on image region selection and automatic network structures determination. We establish the posterior consistency for model parameters and selection consistency for image regions when the number of voxels/pixels grows much faster than the sample size. To ensure computational efficiency, we develop a stochastic gradient Langevin dynamics (SGLD) algorithm for posterior inference. We evaluate our method through extensive comparisons with state-of-the-art deep learning approaches, analyzing multiple real datasets, including task fMRI data from the Adolescent Brain Cognitive Development (ABCD) study.

1. Introduction

Image feature selection in scalar-on-image regression models has recently emerged as a critical topic across various fields, including computer vision and medical imaging. In particular, high-resolution brain imaging data often involve measuring brain signals at hundreds of thousands of pixels or voxels, and these measurements frequently contain substantial noise, creating

[*] Equally contributed

[†] To whom correspondence should be addressed: jiankang@umich.edu

multiple challenges for regression analysis. The high dimensionality of imaging predictors, combined with a low signal-to-noise ratio, makes identifying the true signal difficult, especially given the complex spatial structures inherent in such data. Under these conditions, common regularization methods, such as Lasso, may struggle to deliver optimal performance as they might fail to fully capture these intricate spatial dependencies. Furthermore, the complexity of the system often necessitates nonlinear modeling tools to achieve an accurate model fit.

Deep learning has achieved significant success in many predictive modeling applications (Goodfellow et al., 2016; Hinton et al., 2006; LeCun et al., 2015). Research has shown that deep neural networks can, under certain conditions, mitigate the curse of dimensionality (Bauer and Kohler, 2019) and attain nearly optimal convergence rates for nonparametric regression (Schmidt-Hieber, 2020). The flexibility of neural networks enables them to handle complex nonlinear systems and deliver high prediction accuracy, provided that sufficiently large training datasets are available. In the analysis of image data, convolutional neural networks (CNNs) represent a major breakthrough. However, the outputs of CNNs and other neural networks can be challenging to interpret, and there is no standardized framework for determining the optimal network architecture (Goodfellow et al., 2016). Additionally, deep learning models are prone to overfitting and may produce unreliable results when the sample size is small. This issue is particularly relevant in fields like medical imaging, where obtaining large training datasets is often impractical.

Regularization and feature selection methods are commonly employed in linear models to address the “small data“ problem. Similar strategies have been extended to nonlinear systems. In many studies, deep neural networks have been integrated with sparse regularization techniques to encourage weight sparsity (Liu et al., 2015; Alvarez and Salzmann, 2016; Scardapane et al., 2017; Bach, 2017; Feng and Simon, 2017). Some researchers have also proposed adding specialized layers to perform feature selection. For example, Li et al. (2016b) and Chen et al. (2021) introduced a sparse one-to-one selection layer between the input and subsequent hidden layers to select features. Similarly, Lemhadri et al. (2021) employed a linear skip-layer from the input to the output as a constraint, controlling the sparsity of the weights in the first hidden layer.

In the Bayesian perspective, regularization of the neural network is achieved via priors. For example, the Bayesian regularized neural network (BRNN) used in Okut et al. (2011) and Gianola et al. (2011) applied the Gaussian prior distribution as a regularization term to penalize large connection weights for image feature selection. The parsimonious Bayesian deep network (PBDN) proposed by Zhou (2018) used a gamma process prior to shrink the width and a layer-wise greedy-learning strategy to shrink the depth of the network. Other commonly used priors including the spike-and-slab prior (Polson and Ročková, 2018) and the mixture of Gaussian prior (Sun et al., 2022a) have also been considered to achieve sparsity. Theoretical studies of Bayesian neural networks (BNN) have established the posterior consistency of feature selections, see, e.g., Liang et al. (2018); Sun et al. (2021, 2022b). However, none of these methods are particularly designed for image data and explicitly account for the spatial dependence among the imaging predictors.

For modeling sparse and smooth effects of imaging predictors, a variety of prior specifications have been proposed for scalar-on-image regression, including Ising model, conditionally autoregressive model, and Dirichlet processes (Smith and Fahrmeir, 2007; Goldsmith et al., 2014; Li et al., 2015). In particular, Kang et al. (2018) developed the soft-thresholded

Gaussian process (STGP) for scalar-on-image linear regression for a continuous response and probit regression for a binary response. STGP provides a large probability support over the class of piecewise, sparse and continuous spatially varying functions, leading to the posterior consistency of parameters and region selection in linear regression and probit regression (Kang et al., 2018). However, the linear scalar-on-image regression model (or the probit model) is insufficient to capture the complex association between the response variable and predictors in many applications, which may result in a lack of fit as well as inaccurate region selection and prediction.

In this paper, we propose a novel spatially varying single-layer neural network prior, termed SV-NN, for Bayesian scalar-on-image regression. SV-NN incorporates a single-layer architecture whose weights follow an STGP, thereby promoting both sparsity and piecewise smoothness in the imaging effects. The sparse, spatially varying weights facilitate efficient image region selection, enhancing the model's interpretability. We perform a rigorous theoretical analysis of SV-NN with a single hidden layer by establishing the large support of the prior, posterior consistency of the model parameters, and selection consistency for important image regions. For posterior inference, we develop an efficient MCMC algorithm using the stochastic gradient Langevin dynamics (SGLD, Welling and Teh (2011)). Through extensive empirical evaluations, we demonstrate that SV-NN achieves superior predictive accuracy compared to state-of-the-art deep learning methods across a variety of real-world datasets, particularly when the training sample size is small. We compare SV-NN with STGP, BNN, DNN and CNN via the analysis of the MNIST, Fashion MNIST, and neuroimaging datasets, showing that SV-NN consistently outperforms existing alternatives, especially in small-sample scenarios.

The rest of the paper is organized as follows. Section 2 introduces the model and its prior specifications. Section 3 develops the posterior consistency of the coefficients and the selection consistency of important image regions. Section 4 proposes an efficient posterior computation algorithm using SGLD. In Section 5, we compare our method with existing methods for multiple datasets. Section 6 concludes the paper.

2. Method

In this section, we introduce the framework of the Bayesian scalar-on-image regression and construct the SV-NN prior.

2.1 Bayesian scalar-on-image regression

Suppose $\mathcal{B}$ is a compact subset of the d-dimensional Euclidean space $\mathbb{R}^d$ corresponding to the whole brain region in a neuroimaging study. Let $\mathcal{P} := \{\mathcal{S}_j\}_{j=1}^p$ be a partition of $\mathcal{B}$ such that $\mathcal{S}_{j_1} \cap \mathcal{S}_{j_2} = \emptyset$, $\forall j_1 \neq j_2$ and $\mathcal{B} = \cup_{j=1}^p \mathcal{S}_j$, where each $\mathcal{S}_j$ represents a voxel or a region of the image data and the number of voxels/regions is p, i.e., $|\mathcal{P}| = p$. Suppose we collect n images $\mathbf{X}_i = (X_i(\mathcal{S}_1), \ldots, X_i(\mathcal{S}_p))^\top$, $i = 1, \ldots, n$ on $\mathcal{B}$, where $X_i(\mathcal{S}_j)$ represents the imaging signal on the voxel $\mathcal{S}_j$ for the i-th image. Let $Y_i \in \mathbb{R}$ represent the scalar response variable for the i-th subject. Let $\mathbf{W}_i = (W_{i,1}, \ldots, W_{i,q})^\top \in \mathbb{R}^q$ represent nonimage scalar predictors, and $\mathbf{D}_n = \{Y_i, \mathbf{X}_i, \mathbf{W}_i\}_{i=1}^n$ represent all the data with n observations. We assume that the conditional density of Y_i given $\mathbf{X}_i$ and $\mathbf{W}_i$ belongs to an exponential family, that is,

$$\begin{aligned}&\log\{\pi(Y_i|\mathbf{X}_i, \mathbf{W}_i, \boldsymbol{\alpha}, f)\}\\ &= Y_i\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i)\} - b\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i)\} + A(Y_i, \mathbf{X}_i, \mathbf{W}_i),\end{aligned} \tag{1}$$

where $b(\cdot)$ and $A(\cdot)$ are both known functions that can be chosen according to the data type of Y_i. Let $g(\cdot)$ be a link function such that $g^{-1}(\cdot) = b'(\cdot)$ and $g\{E(Y \mid \mathbf{W}, \mathbf{X})\} = \mathbf{W}^\top\boldsymbol{\alpha} + f(\mathbf{X})$, where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^\top$ captures the linear effects of scalar predictors, and $f : \mathbb{R}^p \to \mathbb{R}$ is an imaging effect function that captures the effects of the imaging predictor.

Our goal is to construct a single hidden layer neural network prior for making inference on the imaging effect function f. We assume that f admits the following representation:

$$f(\mathbf{X}) = \int_{\mathbb{B}} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta})\tau(d\boldsymbol{\beta}), \tag{2}$$

where $\mathbb{B}$ is a compact subset of the Euclidean space $\mathbb{R}^p$, τ is a signed measure with a bounded total variation on $\mathbb{B}$ and $\mathcal{H} : \mathbb{X} \times \mathbb{B} \to \mathbb{R}$ is an image kernel defined as

$$\mathcal{H}(\mathbf{X}, \boldsymbol{\beta}) = h\left\{\sum_{j=1}^{p} \beta(\mathcal{S}_j)X(\mathcal{S}_j)\right\}, \quad \mathbf{X} \in \mathbb{X}, \quad \boldsymbol{\beta} \in \mathbb{B}.$$

Here, $\boldsymbol{\beta} = (\beta(\mathcal{S}_1), \dots, \beta(\mathcal{S}_p))^\top$ represent the effect coefficients, and $\mathbb{X}$ is another compact subset of $\mathbb{R}^p$. The function h is a prespecified activation function satisfying mild conditions, allowing the use of commonly employed activation functions, such as the sigmoid function: $1/\{1 + \exp(-x)\}$, the hyperbolic tangent function: $\tanh(x)$ or the Rectified Linear Unit (ReLU) function (Glorot et al., 2011): $\max\{0, x\}$. In our experiments, we adopt the ReLU function and achieve satisfactory results. The integral representation in (2) has been extensively explored within the neural network literature, particularly in relation to Barron spaces (Barron, 1993). Barron-type spaces are notably broad, encompassing a wide range of high-dimensional functions that can be effectively approximated using shallow neural networks (Petrosyan et al., 2020; Ma et al., 2022). Compared to classical function spaces such as Sobolev and Hölder spaces, Barron-type spaces include a richer class of globally smooth, yet potentially nonlocal, functions (Meng and Ming, 2022; Wojtowytsch et al., 2022). Furthermore, functions within Barron spaces often exhibit dimension-independent approximation rates, contrasting with the dimensionality challenges encountered in traditional smoothness spaces (Mhaskar, 2020). Consequently, the representation in (2) captures a diverse family of functions that extend beyond the classical Sobolev and Hölder classes.

2.2 Prior Specifications

We propose a spatially varying single layer neural network (SV-NN) prior with K hidden units for the unknown function f as follows,

$$\psi(\mathbf{X}; \boldsymbol{\theta}) = K^{-1}\sum_{k=1}^{K} \zeta_k \mathcal{H}(\mathbf{X}, \boldsymbol{\beta}_k), \tag{3}$$

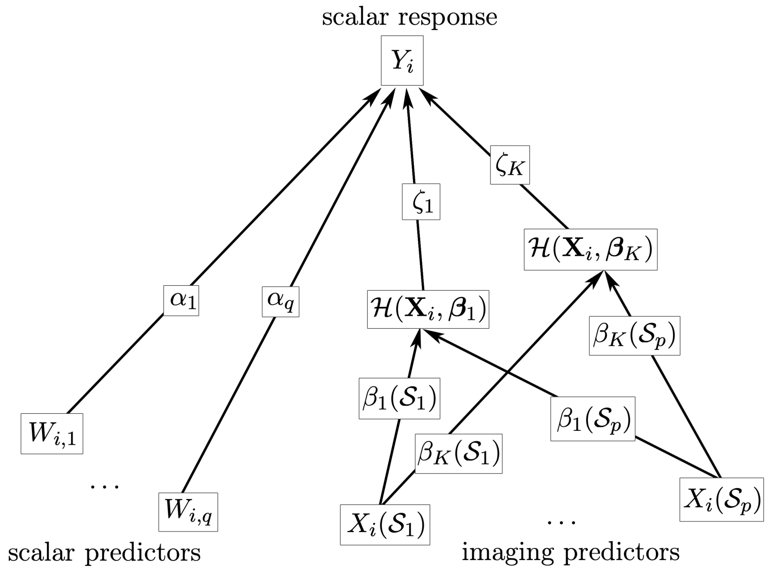


Figure 1: Structure of the Bayesian scalar-on-image regression with SV-NN prior.

where $\boldsymbol{\theta} = \{\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\zeta}\}$ with $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_K)^\top$ denotes the parameter of the network. The weight parameter ζ_k $(k = 1, \ldots, K)$ represents the linear effect of the kth hidden unit on the outcome, for which we assign a Gaussian distribution prior, i.e., $\zeta_k \sim \mathcal{N}(0, \sigma_\zeta^2)$. Each hidden unit is constructed by an image kernel $\mathcal{H}$ and the input imaging features through the unit-specific spatially varying weights $\boldsymbol{\beta}_k = (\beta_k(\mathcal{S}_1), \ldots, \beta_k(\mathcal{S}_p))^\top$, where $\beta_k(\mathcal{S}_j)$ represents the effect of the image feature from $\mathcal{S}_j$ on the hidden unit k. We assume $\beta_k(\cdot)$ is a signed measure and for any $j = 1 \ldots p$, $\beta_k(\mathcal{S}_j)$ has the following form:

$$\beta_k(\mathcal{S}_j) = \int_{\mathcal{S}_j} \phi_k(s) m(ds),$$

where $\phi_k(s)$ represents the k-th imaging signal intensity at spatial location s in the brain region $\mathcal{B}$ and m denotes the Lebesgue measure. We then assign for $\phi_k(\cdot)$ a soft-thresholded Gaussian process (STGP, Kang et al. (2018)) prior, which is constructed by thresholding a Gaussian process $\tilde{\phi}_k(s)$ with a soft-threshold function, i.e.,

$$\phi_k(s) = \mathfrak{T}_\upsilon\{\tilde{\phi}_k(s)\}, \quad \tilde{\phi}_k(s) \sim \mathcal{GP}(0, \kappa), \tag{4}$$

where $\mathfrak{T}_\upsilon(x) = \operatorname{sgn}(x) \max\{0, |x| - \upsilon\}$ with $\upsilon > 0$ is a soft-threshold function; $\operatorname{sgn}(x)$ is the sign of x for any $x \in \mathbb{R}$, i.e. $\operatorname{sgn}(x) = 1$ if $x > 0$, $\operatorname{sgn}(x) = -1$ if $x < 0$ and $\operatorname{sgn}(0) = 0$; and $\mathcal{GP}(0, \kappa)$ is a Gaussian process with zero mean and covariance kernel κ defined as $\kappa(s, s') = \operatorname{Cov}\{\tilde{\phi}_k(s), \tilde{\phi}_k(s')\}$ for any $s, s' \in \mathcal{B}$. We denote such an SV-NN prior as $\mathcal{SV}$-$\mathcal{NN}(K, \sigma_\zeta^2, \upsilon, \kappa)$.

In model (1) with the SV-NN prior, our primary interests focus on the weights $\boldsymbol{\beta}_1, \ldots \boldsymbol{\beta}_K$ and the underlying intensity functions ϕ_k, $k = 1, \ldots, K$. In this prior specification, the sparsity of $\phi_k(s)$ is controlled by the threshold υ; and the spatial smoothness and variance are determined by the covariance kernel κ. For convenience, we represent the covariance kernel $\kappa(s,s') = \sigma_b^2\tilde{\kappa}(s,s')$ where $\sigma_b^2 = \max_{s,s'}\kappa(s,s')$ controls the maximum prior variance of $\phi_k(s)$ and $\tilde{\kappa}(s,s')$ is a re-scaled covariance kernel such that $\tilde{\kappa}(s,s') \le 1$ for any s,s'. Figure 1 illustrates the structure of our proposed model, for which we perform rigorous theoretical analyses in Section 3. Note that we can construct additional numbers of fully connected hidden layers upon the first layer and construct a deep neural network with spatially varying weights at the input layer. In our numerical experiments in Section 5, we have tested this architecture and obtained promising results that validate the efficacy of neural networks with more than one layer.

3. Theory

In this section, we present the theoretical properties of the proposed scale-on-image regression model with the SV-NN prior. Our analysis covers the large support of the prior, the posterior consistency of the imaging operator, and the consistency of region selection. Detailed proofs of these results are provided in Appendix A.

We first introduce additional notation. Let $\|\phi\|_q = \left\{\int|\phi(\xi)|^q m(d\xi)\right\}^{1/q}, q \ge 1$ and $\|\phi\|_\infty = \sup|\phi(\xi)|$ be the L^q-norm and the L^∞-norm of a real function ϕ, with respect to the Lebesgue measure $m(.)$, respectively; let $\|\mathbf{v}\|_q = \left(\sum_{i=1}^d |v_i|^q\right)^{1/q}, q \ge 1$ be the L^q-norm and $\|\mathbf{v}\|_\infty = \max_{i=1}(|v_i|)$ be the L^∞-norm for any vector $\mathbf{v} \in \mathbb{R}^d$. Let $\phi \in \mathcal{C}^\rho(\mathcal{D})$ denote a differentiable function of order ρ on $\mathcal{D}$. Let $O_\delta(\mathbf{b}) := \{\mathbf{z} \in \mathbb{B}, \|\mathbf{z}-\mathbf{b}\|_1 < \delta\}$ denote the δ-neighborhood of $\mathbf{b}$. Fox simplicity, we assume $\boldsymbol{\alpha}$ is known in this section as our major interest focuses on the spatially varying weights. The unknown parameter become only $\boldsymbol{\theta} = \{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K,\boldsymbol{\zeta}\}$. Denote by f^* the true imaging effect function that generates data $\mathbf{D}_n$ in model (1), and $\mathcal{F}$ the functional space, which satisfies the following condition.

Condition 1 *The true imaging effect function f^* belongs to a functional space $\mathcal{F}$ such that for any $f \in \mathcal{F}$ and partition $\mathcal{P}$ of $\mathcal{B}$, $\mathcal{P} := \{\mathcal{S}_j\}_{j=1}^p$, we have*

$$f(\mathbf{X}) = \int_{\mathbb{B}} \mathcal{H}(\mathbf{X},\boldsymbol{\beta})\tau(d\boldsymbol{\beta}),$$

where $\boldsymbol{\beta} = \{\beta(\mathcal{S}_1),\ldots,\beta(\mathcal{S}_p)\}^\top \in \mathbb{B}$ is a p-dimensional vector, β is a signed measure, $\mathcal{H}$ is the image kernel, and τ is a r-admissible signed measure, i.e., $|\tau|(\mathbb{B}) < M_b < \infty$, the support supp($\tau$) is a r-dimensional subset of $\mathbb{B}$, and there exists a constant C such that $|\tau|(O_\delta(\mathbf{b})) \le C\delta^r|\tau|(\mathbb{B})$, $\forall \mathbf{b} \in \mathbb{B}, 0 < \delta \le 1$.

Condition 2 *For any $\mathcal{R} \in \mathcal{B}, \beta(\mathcal{R}) = \int_{\mathcal{R}} \phi(s)m(ds), \phi \in \mathcal{I}$. For any $\phi \in \mathcal{I}$, there exist open sets $\mathcal{R}_1$ and $\mathcal{R}_{-1}$ with $\bar{\mathcal{R}}_1 \cap \bar{\mathcal{R}}_{-1} = \emptyset$, $\bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1} \neq \emptyset$, and a) ϕ is smooth over $\bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1}$, i.e., $\phi(s)I(s \in \bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1}) \in \mathcal{C}^{\lceil d/2\rceil}(\bar{\mathcal{R}}_1 \cup \bar{\mathcal{R}}_{-1})$; b) $\phi(s) = 0$ for $s \in \mathcal{R}_0$, $\phi(s) > 0$ for $s \in \mathcal{R}_1$ and $\phi(s) < 0$ for $s \in \mathcal{R}_{-1}$, where $\mathcal{R}_0 = \mathcal{B} - (\mathcal{R}_1 \cup \mathcal{R}_{-1})$; c) ϕ is continuous over $\mathcal{B}$.*

In Condition 1, we assume that the true imaging effect function is determined by a partition of the image space $\mathcal{P}$, a nonlinear image kernel $\mathcal{H}$, and a signed measure τ defined in the vector space $\mathbb{B}$ of the effect coefficients $\boldsymbol{\beta}$. Condition 2 further assumes the effect coefficients are driven by an underlying signal intensity function ϕ, which is sparse, piecewise smooth and continuous, allowing it to effectively capture important regions from the imaging predictors. Condition 3 specifies both the non-linear image kernel $\mathcal{H}$ and the link function g^{-1} used in model (1). The partition $\mathcal{P}$ corresponds to the resolution of images, which will be further elaborated in Condition 5.

Condition 3 *The image kernel has the expression* $\mathcal{H}(\mathbf{X}, \boldsymbol{\beta}) = h\{\sum_{j=1}^{p} \beta(\mathcal{S}_j)X(\mathcal{S}_j)\}$*, and* $h(t)$ *is a Lipschitz continuous non-linear activation function that satisfies with* $|h(t) - h(t')| < M_0|t - t'|$*, where* M_0 *is a constant. The link* $g^{-1} = b'$ *is a continuously differentiable and strictly monotonic function.*

Condition 3 assumes the activation function is Lipschitz continuous. Similar conditions have been applied in the literature, see, e.g., Liang et al. (2018) and Sun and Liang (2022). Commonly used activation functions including $\tanh(\cdot)$, the sigmoid function, and the ReLU function satisfies this assumption. Additionally, Condition 3 assumes the link function is continuously differentiable and monotonic, making it applicable to a wide range of models.

3.1 Large Support

In this section, we discuss the large support of the proposed SV-NN prior (3). We first construct a series of sieves for the imaging effect function, and then show that we can find a single layer neural network as a good approximation with dimension independence approximation bound on the sieves. The sieves are defined as

$$\mathcal{F}_p = \left\{ f \in \mathcal{F} : f(\mathbf{X}) = \int_{\mathbb{B}_p} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta}) \tau(d\boldsymbol{\beta}) \right\}$$

with

$$\mathbb{B}_p := \left\{ \boldsymbol{\beta} \in \mathbb{B} : \beta(\mathcal{R}) = \int_{\mathcal{R}} \phi(s) m(ds), \|\phi\|_\infty \leq p^{1/2}, \forall \mathcal{R} \in \mathcal{B} \right\}.$$

In addition, we need the following Condition 4 for the observed imaging predictors, and Condition 5 for the partition of the image space. Condition 4 ensures that the total information that we can extract from imaging predictors (measured with L^2-norm) keeps bounded as the resolution of image goes to infinity. Condition 5 assumes that each voxel or subregion of the image has a comparable size. These conditions are natural assumptions in applications and have been applied in the previous study of neuroimaging models (Wu et al., 2024).

Condition 4 *The imaging predictors* $\mathbf{X}_i := (X_i(\mathcal{S}_1), \ldots, X_i(\mathcal{S}_p))^\top$*,* $i = 1, \ldots, n$ *belong to a compact set* $\mathbb{X} \subseteq \mathbb{R}^p$ *such that for any* $\mathbf{X} \in \mathbb{X}$*,* $\|\mathbf{X}\|_2 \leq M_1$*, where* $M_1 > 0$ *is a constant.* $\mathbf{X}_i$ *are independent samples from a distribution* Q *on* $\mathbb{X}$*, and* Q *has a Lebesgue density* π_x *bounded below by a positive constant. Given* $\mathbf{X}_i$*,* $\mathbf{W}_i$ *and the true model,* $E_{f^*}\left(Y_i^2 \mid \mathbf{X}_i, \mathbf{W}_i\right) < M_1$*.*

Condition 5 *The partition* $\mathcal{P} := \{\mathcal{S}_j\}_{j=1}^{p}$ *of* $\mathcal{B}$ *satisfies that: a)* $\forall j$*,* $\mathcal{S}_j \neq \emptyset$ *and* $m(\mathcal{S}_j) \leq \bar{m}(\mathcal{S}_j) < \infty$*, where* $m(.)$ *is the Lebesgue measure and* $\bar{m}(\mathcal{S}_j) := \sup_{s_1, s_2 \in \mathcal{S}_j} (\|s_1 - s_2\|_\infty)^d$*; b) There exists a constant* $0 < L < m(\mathcal{B})$ *such that* $\max_{j=1,\ldots,p} \bar{m}(\mathcal{S}_j) < (Lp)^{-1}$*, as* $p \to \infty$*.*

Now we can show with the following Lemma 1 that a neural network with a single hidden layer architecture as in equation (3) can be a good approximation of the imaging effect function with a dimension independent approximation bound.

Lemma 1 *Suppose Conditions 1 - 5 hold. For any function $f \in \mathcal{F}_p$, there exist an integer C such that for sufficiently large K and $\mathbf{X} \in \mathbb{X}$, we have*

$$|\psi(\mathbf{X};\boldsymbol{\theta}) - f(\mathbf{X})| \leq CM_b \left(\frac{\log K}{K}\right)^{1/2} K^{-1/r},$$

where $\psi(\mathbf{X};\boldsymbol{\theta}) = K^{-1}\sum_{k=1}^{K} \zeta_k \mathcal{H}(\mathbf{X},\boldsymbol{\beta}_k)$ with $|\zeta_k| < M_b$ and $\boldsymbol{\beta}_k \in \mathbb{B}_p$, $k = 1,\ldots,K$.

To show the SV-NN prior has large support on the functional space of interest, we also need the following Condition 6 for the smoothness of the covariance kernel κ of the soft-thresholded Gaussian process.

Condition 6 *For any fixed s, the covariance kernel $\kappa(s,.)$ has continuous partial derivatives up to order $2\lceil d/2\rceil + 2$.*

Now we establish the large support property of the SV-NN prior with Theorem 1.

Theorem 1 *Suppose Conditions 1 - 6 hold. For any $f^* \in \mathcal{F}$ and $\varepsilon > 0$, there exist sufficiently large p and K, such that the SV-NN prior $f \sim \mathcal{SV}\text{-}\mathcal{NN}(K, \sigma_\zeta^2, \upsilon, \kappa)$ satisfies*

$$\Pr\left(\|f - f^*\|_\infty < \varepsilon\right) > 0.$$

The large support property ensures a positive prior probability such that the imaging effect function f falls into an arbitrarily small neighborhood of the true $f^* \in \mathcal{F}$, which is assumed to be driven by an underlying smooth and sparse imaging signal intensity function. Therefore, we can conceptually make the image region selection with the SV-NN prior.

3.2 Posterior Consistency

Applying the general Bayesian consistency theory (Choudhuri et al., 2004), we establish the posterior consistency by verifying two conditions: the prior positivity of the neighborhoods in Lemma 2 and the existence of uniformly consistent tests in Lemmas 3 - 6.

Lemma 2 *Suppose Conditions 1 - 6 hold. Denote by $\pi_n(.;f)$ the conditional density function of $\mathbf{D}_{n,i} = \{Y_i, \mathbf{X}_i, \mathbf{W}_i\}$ given f. Let $\Lambda_n(.;f^*,f) = \log \pi_n(.;f^*) - \log \pi_n(.;f)$. Define*

$$K_{n,i}(f^*,f) = E_{f^*}\left\{\Lambda_n(\mathbf{D}_{n,i};f^*,f)\right\} \text{ and } V_{n,i}(f^*,f) = \mathrm{Var}_{f^*}\left\{\Lambda_n(\mathbf{D}_{n,i};f^*,f)\right\}.$$

There exists a set $\tilde{\mathcal{F}}$ with $\Pr(\tilde{\mathcal{F}}) > 0$ such that for any $\varepsilon > 0$,

$$\liminf_{n\to\infty} \Pr\left(f \in \tilde{\mathcal{F}}, \frac{1}{n}\sum_{i=1}^{n} K_{n,i}(f^*,f) < \varepsilon\right) > 0,$$

$$\lim_{n\to\infty} \frac{1}{n^2}\sum_{i=1}^{n} V_{n,i}(f^*,f) = 0, \ \forall\ f \in \tilde{\mathcal{F}}.$$

We then need to construct a series of uniformly consistent tests on the sieves, which requires a detailed specification of the order of sample size n, the number of voxels p, and the number of hidden units K in the neural network.

Condition 7 *There exist $L_0 > 0$, $L_1 > 0$, $N > 0$ and $0 < \nu < 1$ such that for all $n > N$, $L_0 K^2 < n^{1-\nu}$ and $n < L_1 p$.*

Condition 7 implies that the order of p grows polynomially with n, which is consistent with many neuroimaging studies in which the number of voxels is much larger than the number of images. Furthermore, Condition 7 suggests that a relatively small neural network, with the number of hidden units K smaller than both n and p, is sufficient for the SV-NN prior. Under Conditions 1 - 7, we demonstrate through Lemmas 3 and 4 that the sieves defined above exhibit the desired properties.

Lemma 3 *Under Conditions 1 - 7, the ε-covering number $N(\varepsilon, \mathcal{F}_p, \|.\|_\infty)$ of $\mathcal{F}_p$ in the supremum norm satisfies*

$$\log N(\varepsilon, \mathcal{F}_p, \|.\|_\infty) \leq C_1 n^{1-\nu/2} \varepsilon^{-2},$$

where $C_1 > 0$ is a constant.

Lemma 4 *If $f \sim \mathcal{SV}\text{-}\mathcal{NN}(K, \sigma_\zeta^2, \upsilon, \kappa)$ and Conditions 1 - 7 hold, there exist constants C_0 and C_1 such that for all $n \geq 1$, $\Pr(\mathcal{F}_p^C) \leq C_0 \exp(-C_1 n)$.*

Now on each $\mathcal{F}_p$, the key step is to construct an efficient test Ψ_n to distinguish the true function f^* from any function f that is sufficiently different from f^*. For a scalar-on-image linear regression model Kang et al. (2018), the linearity of the model simplifies the construction procedure. Since we propose a novel nonlinear regression framework with a general imaging effect function, we require Lemma 5 to ensure that the model remains identifiable under the nonlinear transformation.

Lemma 5 *Under Conditions 1 - 5 and 7, for any $\varepsilon > 0$, $0 < \epsilon < \varepsilon^2$, and $f, f^* \in \mathcal{F}_p$, let*

$$\mathcal{A}_n = \left\{ \sum_{i=1}^{n} \left| g^{-1}\left\{ \mathbf{W}^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i) \right\} - g^{-1}\left\{ \mathbf{W}^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} \right| \geq n\epsilon \right\}.$$

If $\|f - f^*\|_1 > \varepsilon$, then there exists a constant C_0 such that $\Pr\left(\mathcal{A}_n^C\right) \leq \exp(-C_0 n)$ and $\Pr\left(\cup_{m=1}^\infty \cap_{n=m}^\infty \mathcal{A}_m\right) = 1$.

Lemma 5 implies that model (1) is identifiable with probability one when the imaging predictors are realizations of a random variable under some regularity conditions. This result provides a foundation for the construction of the uniformly consistent tests in Lemma 6.

Lemma 6 *Under Conditions 1 - 7, for any $\varepsilon > 0$ there exist $N, C > 0$ such that for all $n > N$ and all $f \in \mathcal{F}$, if $\|f - f^*\|_1 > \varepsilon$, there exists a test function Ψ_n such that $E_{f^*}(\Psi_n) < \exp(-Cn)$ and $E_f(1 - \Psi_n) \leq \exp(-Cn)$.*

Lemmas 3-6 verify the conditions of constructing uniformly consistent tests required in Choudhuri et al. (2004), based on which we establish the posterior consistency result with Theorem 2, and further the posterior selection consistency of important image regions with Theorem 3. Let $P_{f^*}^{(n)}$ be the actual distribution of data $\mathbf{D}_n$ given the true function $f^* \in \mathcal{F}$.

Theorem 2 *Under Conditions 1 - 7, for any $f^* \in \mathcal{F}$ and $\varepsilon > 0$, as $n, p, K \to \infty$,*

$$\Pr(f \in \mathcal{F} : \|f - f^*\|_1 < \varepsilon \mid \mathbf{D}_n) \to 1$$

in $P_{f^*}^{(n)}$ probability.

Theorem 2 establishes the posterior consistency of the imaging effect function f, which ensures the posterior distribution of f concentrates on the ε-neighborhood of the true imaging effect function for any $f^* \in \mathcal{F}$ and $\varepsilon > 0$, as the sample size n, the number of regions p and the number of hidden units K of the network go to infinity. For simplicity, we present the posterior consistency in terms of the Lebesgue L^1 distance. If we relax the requirement for the existence of Lebesgue density of Q in Condition 4, we can similarly establish consistency in the L^1 distance with respect to the Q probability measure.

3.3 Selection Consistency and Screening Property of Important Regions

To investigate the region importance of the model, we define the effect variation on the jth region as

$$\Delta_j|f| := \sup_{0<|\delta|<c} \int_{\mathbb{X}_\delta} |f(\mathbf{X}) - f(\mathbf{X} + \delta \mathbf{e}_j)| m(d\mathbf{X}), \tag{5}$$

where $\mathbf{e}_j = (e_1, \dots, e_p)^\top$ with $e_j = 1$, $e_{j'} = 0$, $j' \neq j$, δ is a constant which measures the signal variation of the input image on the jth region, c is a upper bound of the variation we consider, and $\mathbb{X}_\delta := \{\mathbf{X} \in \mathbb{X} : \mathbf{X} + \delta \mathbf{e}_j \in \mathbb{X}\}$ is a subspace of $\mathbb{X}$, which ensures the variation is restricted to a meaningful range. A region can be regarded as less important if $\Delta_j|f| < \varepsilon$ for any c, where ε measures the minimum detectable effect variation. In contrast, a region is important if a small c leads to a large effect variation $\Delta_j|f|$. Theorem 3 indicates that selection of the important regions also achieves posterior consistency.

Theorem 3 *Under Conditions 1 - 7, for any $f^* \in \mathcal{F}$, $\delta > 0$ and $\varepsilon > 0$, as $n, p, K \to \infty$,*

$$\Pr\left(\sup_j |(\Delta_j|f| - \Delta_j|f^*|)| < \varepsilon \mid \mathbf{D}_n\right) \to 1,$$

in $P_{f^*}^{(n)}$ probability.

Furthermore, we present the screening property of important regions with the following Corollary 4. This property ensures we can screen out unimportant regions with evaluating the weights $\boldsymbol{\beta}_k$, $k = 1, \dots, K$, rather than the imaging effect function f, which reduces computational burden during the procedure of posterior inferences. The details of the posterior inferences are discussed in the next section.

Corollary 4 *Define index sets*

$$\mathcal{I}_0 := \left\{ j : \max_{k=1,\ldots,K} |\beta_k(\mathcal{S}_j)| = 0 \right\} \text{ and } \mathcal{I}_\varepsilon^* := \{ j : \Delta_j |f^*| < \varepsilon \}.$$

Then for any $f^* \in \mathcal{F}$ and $\varepsilon > 0$, under Conditions 1 - 7, as $n, p, K \to \infty$, we have

$$\Pr(\mathcal{I}_0 \subseteq \mathcal{I}_\varepsilon^* \mid \mathbf{D}_n) \to 1$$

in $P_{f^*}^{(n)}$ probability.

4. Posterior Computation

In this section, we develop an equivalent representation of the SV-NN prior to facilitate efficient posterior computation. The model representation is general and valid for any covariance kernel κ. In particular, we discuss a special covariance kernel and its closed form of the eigen decomposition, which can be used for efficient computation in large scale imaging data analysis. Finally, we present the details of the stochastic gradient MCMC algorithm.

4.1 Model Representation

By Mercer's theorem (Williams and Rasmussen, 2006), under some mild regularity conditions, the covariance kernel $\tilde{\kappa}(s, s')$ can be decomposed as $\tilde{\kappa}(s, s') = \sum_{l=1}^\infty \lambda_l \varphi_l(s) \varphi_l(s')$, where $\{\lambda_l\}_{l=1}^\infty$ are eigen values with $\lambda_l \geq \lambda_{l+1} > 0$ for $l \geq 1$ and $\{\varphi_l(s)\}_{l=1}^\infty$ are orthonormal eigen functions which satisfy the following properties: $\int_{\mathbb{R}^d} \varphi_l^2(s) ds = 1$, for $l \geq 1$ and $\int_{\mathbb{R}^d} \varphi_l(s) \varphi_{l'}(s) ds = 0$, for $l \neq l'$. Then we represent the GP $\tilde{\phi}_k(s)$ in (4) using the Karhunen-Loève expansion and obtain an equivalent representation of the STGP prior for $\phi_k(s)$ by taking $\tilde{\upsilon} = \upsilon / \sigma_b$. Specifically, for $k = 1, \ldots, K$ and $s \in \mathcal{B}$,

$$\phi_k(s) = \sigma_b \mathfrak{T}_{\tilde{\upsilon}} \left\{ \sum_{l=1}^\infty b_{kl} \varphi_l(s) \right\}, \quad b_{kl} \sim \mathcal{N}(0, \lambda_l), \tag{6}$$

where the eigen functions $\{\varphi_l(s)\}_{l=1}^\infty$ and the eigen values $\{\lambda_l\}_{l=1}^\infty$ are determined by the covariance kernel function $\tilde{\kappa}(s, s')$. Coefficients $\{b_{kl}\}_{l=1}^\infty$ follow *a priori* independent Gaussian distributions, but involve an infinite number of parameters. In practice, we approximate (6) by truncating the expansion at a finite number, say L, of basis functions. Let $\mathbf{b}_k = (b_{k1}, \ldots, b_{kL})^\top$, $\boldsymbol{\varphi}(s) = (\varphi_1(s), \ldots, \varphi_L(s))^\top$ and $\mathbf{\Lambda} = \text{diag}\{\lambda_1, \ldots, \lambda_L\}$. Then the approximation of the SV-NN prior can be represented as

$$\begin{aligned} \psi(\mathbf{X}; \boldsymbol{\theta}) &= K^{-1} \sum_{k=1}^K \zeta_k \mathcal{H}(\mathbf{X}, \boldsymbol{\beta}_k), \quad \zeta_k \sim \mathcal{N}(0, \sigma_\zeta^2), \\ \beta_k(\mathcal{S}_j) &= \phi_k(s_j) m(\mathcal{S}_j), \\ \phi_k(s_j) &= \sigma_b \mathfrak{T}_{\tilde{\upsilon}} \left\{ \mathbf{b}_k^\top \boldsymbol{\varphi}(s_j) \right\}, \quad \mathbf{b}_k \sim \mathcal{N}(0, \mathbf{\Lambda}), \end{aligned} \tag{7}$$

where s_j is the centroid of the region $\mathcal{S}_j$. This prior approximation is valid for any covariance kernel $\tilde{\kappa}(s, s')$, where the corresponding truncated eigen functions $\boldsymbol{\varphi}(s)$ and eigen values $\mathbf{\Lambda}$ can be obtained by numerical approximations using the eigen decomposition of the gram matrix (Williams and Rasmussen, 2006, Chapter 4.3). In the next section, we discuss one special covariance kernel for posterior computation in practice.

4.2 Modified Squared Exponential Kernel

For all the numerical experiments in this paper, we consider a special covariance kernel: the modified squared exponential (MSE) kernel, which is defined as

$$\tilde{\kappa}(s, s') = \exp\{-a(\|s\|_2^2 + \|s'\|_2^2) - b\|s - s'\|_2^2\}, \text{ for } a > 0, b > 0.$$

The MSE kernel becomes the squared exponential (SE) kernel when $a = 0$. For the latent process $\tilde{\phi}_k \sim \mathcal{GP}(0, \sigma_b^2 \tilde{\kappa})$, we have $\mathrm{Var}\{\tilde{\phi}_k(0)\} = \sigma_b^2$, $\mathrm{Var}\{\tilde{\phi}_k(s)\} = \sigma_b^2 \exp(-a\|s\|_2^2)$ and $\mathrm{Cor}\{\tilde{\phi}_k(s), \tilde{\phi}_k(s')\} = \exp\{-b\|s - s'\|_2^2\}$. Thus, the maximum marginal variance of $\tilde{\phi}_k(s)$ is achieved in the center 0. The parameter a controls the decay rate of $\mathrm{Var}\{\tilde{\phi}_k(s)\}$ compared to $\mathrm{Var}\{\tilde{\phi}_k(0)\}$. A smaller a corresponds to a slower decay rate. The parameter b controls the smoothness of $\phi_k(s)$. A smaller b corresponds to a smoother $\tilde{\phi}_k(s)$. As the key attractive property, the MSE kernel has a closed form of the eigen decomposition, where the eigen functions can be represented as Hermite polynomials. Specifically, let $H_k(x) = (2^k k! \sqrt{\pi})^{-1/2} (-1)^k \exp(x^2) \frac{d^k}{dx^k} \exp(-x^2)$ be the kth order normalized Hermit polynomial. For each $m = 0, 1, \ldots,$ and each $l = \binom{m+d-1}{d} + 1, \ldots, \binom{m+d}{d}$, we can find a unique set of non-negative integers $\{m_u\}_{u=1}^d$ such that $\sum_{u=1}^d m_u = m$ and

$$\varphi_l(s) = (2c)^{d/4} \exp(-c\|s\|_2^2) \prod_{u=1}^{d} H_{m_u}(\sqrt{2c} s_u), \text{ for any } s = (s_1, \ldots, s_d)^\top,$$

with $\lambda_l = (\pi/A)^d B^m$, where $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$ and $B = b/A$. Note that we define $\binom{d-1}{d} = 0$ for $d \geq 1$. The integer m is the polynomial degree of $\varphi_l(s)$ and the corresponding eigen values λ_l are the same for the same m. In the truncated eigen decomposition, the total number of eigen functions L can be determined by the maximum polynomial degrees, say M, i.e. $L = \binom{M+d}{d}$. In our experiments where $\mathcal{B} = [-1, 1]^d$, $a = 0.01$ and $b = 10$, we choose the maximum polynomial degree $M = 20$ which corresponds to $L = 231$ for $d = 2$ and $L = 1771$ for $d = 3$.

4.3 Markov chain Monte Carlo

Traditional Markov chain Monte Carlo (MCMC) algorithms such as the Metropolis-Hastings algorithm and Gibbs sampling can be inefficient for posterior computation for Bayesian neural networks due to the high computational costs. Stochastic versions of gradient-based MCMC algorithms serve as scalable variants for complex models and large datasets, among which the stochastic gradient Langevin dynamics (SGLD) algorithm (Welling and Teh, 2011) is the simplest to implement and has a reasonable runtime. It relates stochastic optimization to a first-order Langevin dynamic MCMC by adding a noise term to stochastic gradient descent iterating, and thus skips the Metropolis-Hastings accept-reject step for reducing the computational cost.

Let $\tilde{\boldsymbol{\theta}} = \{\boldsymbol{\zeta}, \boldsymbol{\alpha}, \mathbf{b}, \sigma_b\}$ with $\mathbf{b} = \{\mathbf{b}_k\}_{k=1}^K$ be the parameters of the approximate representation of the model. Algorithm 1 describes the posterior sampling procedure based on SGLD. The input includes training data $\mathbf{D}_n$, number of hidden units K, soft threshold parameter $\tilde{v}$ and hyperparameters a, b, L in the GP decomposition. In each iteration, SGLD first randomly selects a minibatch of the whole data set and approximates the true gradient.

Algorithm 1 Posterior sampling of Bayesian scalar-on-image regression with the SV-NN prior

```
Input: $\mathbf{D}_n = \{Y_i, \mathbf{X}_i, \mathbf{W}_i\}_{i=1}^n$, $K, a, b, \tilde{\upsilon}, L$,
       step size $\{\epsilon_t\}$, minibatch size $|\mathcal{J}|$, number of epochs $T$,
       maximum saves $n_{\max}$, burn-in time $t_{\text{burn}}$.
Output: MCMC samples of parameters
procedure SGLD
    Compute the approximate eigen decomposition $\boldsymbol{\varphi}(s)$ and $\boldsymbol{\Lambda}$ of the MSE kernel $\tilde{\kappa}(s, s')$ given $a$, $b$ and $L$.
    for $t = 1, \cdots, T$ do
        Randomly split sample indices into $J$ disjoint subsets: $\{i\}_{i=1}^n = \bigcup_{j=1}^J \mathcal{J}_j$.
        for $j = 1, \cdots, J$ do
            $\nabla \hat{U}_{\mathcal{J}_j}(\tilde{\boldsymbol{\theta}}) \leftarrow \frac{n}{|\mathcal{J}_j|} \sum_{i \in \mathcal{J}_j} \nabla \log \pi(Y_i | \mathbf{X}_i, \mathbf{W}_i, \tilde{\boldsymbol{\theta}}) + \nabla \log \pi(\tilde{\boldsymbol{\theta}})$
            $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}} + \epsilon_t \nabla \hat{U}_{\mathcal{J}_j}(\tilde{\boldsymbol{\theta}}) + \mathcal{N}(0, 2\epsilon_t)$
        if $t > t_{\text{burn}}$ then
            if number of current saves $> n_{\max}$ then
                Delete the oldest save of $\tilde{\boldsymbol{\theta}}$.
            Save current value of $\tilde{\boldsymbol{\theta}}$ as one posterior sample.
```

Let $\mathcal{J} \subseteq \{1, \ldots, n\}$ be a random subset of the whole sample indices and the subset based gradient is defined as

$$\nabla \hat{U}_{\mathcal{J}}(\tilde{\boldsymbol{\theta}}) = \frac{n}{|\mathcal{J}|} \sum_{i \in \mathcal{J}} \nabla \log \pi(Y_i \mid \mathbf{X}_i, \mathbf{W}_i, \tilde{\boldsymbol{\theta}}) + \nabla \log \pi(\tilde{\boldsymbol{\theta}}),$$

where $\pi(Y_i \mid \mathbf{X}_i, \mathbf{W}_i, \tilde{\boldsymbol{\theta}})$ is the likelihood function of $\tilde{\boldsymbol{\theta}}$ specified according to the sampling distribution of data in (1) and the joint prior distribution $\pi(\tilde{\boldsymbol{\theta}})$ is specified according to SV-NN approximation in (7). The computation of gradient completely follows that of a general neural network, except that we need an approximation to the derivative of the soft-thresholding function. i.e. $\partial \mathfrak{T}_{\tilde{\upsilon}}(x)/\partial x = 1$ if $|x| \geq \tilde{\upsilon}$ and $\partial \mathfrak{T}_{\tilde{\upsilon}}(x)/\partial x = 0$, otherwise. Based on the approximated $\nabla \hat{U}_{\mathcal{J}}(\tilde{\boldsymbol{\theta}})$, SGLD updates the parameter at $t + 1$th iterations as follows:

$$\tilde{\boldsymbol{\theta}}_{t+1} = \tilde{\boldsymbol{\theta}}_t + \epsilon_t \nabla \hat{U}_{\mathcal{J}}(\tilde{\boldsymbol{\theta}}_t) + \mathcal{N}(0, 2\epsilon_t).$$

Here $\{\epsilon_t\}$ is a sequence of step sizes. Theoretically, ϵ_t should decay to zero at an appropriate speed (e.g. polynomial decay $\epsilon_t = a_0(b_0 + t)^{-\gamma_0}$) to ensure convergence. In practice, for simplicity, we use a constant step size. In addition to SGLD, stochastic gradient MCMC (SG-MCMC) algorithms such as Stochastic Gradient Hamiltonian Monte Carlo (SGHMC) (Chen et al., 2014) and preconditioned SGLD (p-SGLD) (Li et al., 2016a) can also be used to sample from the posterior distribution. However, they need additional effort to tune hyperparameters and may suffer from a higher computational burden. Other sampling methods designed for neural networks, such as the Laplace approximation (Ritter et al., 2018) and variational Bayes (Blei et al., 2017) approximate the posterior with a simpler tractable distribution, but the error between their posterior samples and the true posterior tends to be larger than SG-MCMC due to this trade-off.

4.4 Bayesian Inference

We perform image region selection and model prediction within a fully Bayesian inference framework. Theorem 3 guarantees that our model achieves region selection consistency based on the effect variation defined in equation (5). Furthermore, Corollary 4 suggests that we can preliminarily screen out unimportant regions by evaluating the posterior samples of $\boldsymbol{\beta}_k$. Building on this insight, we propose a two-stage thresholding inference procedure for selecting important regions in this section.

Let $\mathfrak{p}_j = \Pr\{\max_k |\beta_k(\mathcal{S}_j)| \neq 0 \mid \mathbf{D}_n\}$ represent the posterior probability that at least one of the spatially varying weights $\beta_k(\mathcal{S}_j)$ of the K hidden units are nonzero at region $\mathcal{S}_j$, $j = 1, \ldots, p$. We refer to $\mathfrak{p}_j$ as the spatially varying posterior inclusion probability (PIP) for the region $\mathcal{S}_j$. Suppose the proposed MCMC algorithm generates T samples of $\boldsymbol{\beta}_k$ and ζ_k after burn-in, denoted as $\{\boldsymbol{\beta}_k^{(t)}, \zeta_k^{(t)}\}_{t=1}^T$. We estimate the PIPs as $\hat{\mathfrak{p}}_j = T^{-1}\sum_{t=1}^T I\{\max_k |\beta_k^{(t)}(\mathcal{S}_j)| \neq 0\}$, where $I(\mathcal{A}) = 1$ if event $\mathcal{A}$ occurs, $I(\mathcal{A}) = 0$ otherwise. To select important image regions, we threshold the estimated PIPs $\{\hat{\mathfrak{p}}_j\}_{j=1}^p$ by controlling the expected Bayesian false discovery rate (FDR) (Morris et al., 2008) at a prespecified level. Specifically, let $\hat{\mathfrak{p}}_{(j)}$ be the jth largest estimated PIP among $\{\hat{\mathfrak{p}}_j\}_{j=1}^p$, i.e. $\hat{\mathfrak{p}}_{(1)} \geq \ldots \geq \hat{\mathfrak{p}}_{(p)}$. We define the first-stage region selection indicators as

$$\hat{\delta}_j = I\{\hat{\mathfrak{p}}_j \geq \hat{\mathfrak{p}}_{(j_{\alpha_1})}\} \text{ with } j_{\alpha_1} = \max\left[l : l^{-1}\sum_{j=1}^{l}\left\{1 - \hat{\mathfrak{p}}_{(j)}\right\} \leq \alpha_1\right].$$

Then, based on $\hat{\delta}_j$ we obtain an index set $\mathcal{J}_1 := \{j : \hat{\delta}_j = 1\}$ of candidates of important regions. The screening property in Corollary 4 ensures that $\mathcal{J}_1$ contains the true set of important regions as a superset. Furthermore, we can define the second-stage posterior inclusion probabilities (PIPs) and region selection indicators by evaluating the effect variations using equation (5) in a similar manner. This approach offers a more refined approximation of the true set of important regions. However, we observe that the first-stage strategy is practically effective in identifying important image regions and often yields similar selection results to the second-stage indicators. Therefore, for the sake of computational efficiency, we report the results obtained using the first-stage strategy in the experiments below. For prediction, we use the posterior predictive mean of the response variable. Suppose that we obtain T posterior samples of parameters $\tilde{\boldsymbol{\theta}}$ given the training data $\mathbf{D}_n$, denoted as $\{\tilde{\boldsymbol{\theta}}^{(t)}\}_{t=1}^T$. Given the test image $\mathbf{X}^*$ and the covariate $\mathbf{W}^*$, the posterior predictive mean of Y^* is $E(Y^* \mid \mathbf{X}^*, \mathbf{W}^*) = \int E(Y^* \mid \mathbf{X}^*, \mathbf{W}^*, \tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}} \mid \mathbf{D}_n)d\tilde{\boldsymbol{\theta}}$ which can be estimated by $T^{-1}\sum_{t=1}^T E(Y^* \mid \mathbf{X}^*, \mathbf{W}^*, \tilde{\boldsymbol{\theta}}^{(t)})$.

5. Experiments

In this section, we evaluate the performance of SV-NN on multiple datasets, including synthetic nonlinear classification datasets, MNIST, Fashion MNIST (Xiao et al., 2017) and neuroimaging data from the Adolescent Brain Cognitive Development study (Casey et al., 2018). The two main evaluation criteria are prediction accuracy and stability of region selection. For each dataset, we consider U random splits on the training and test data. Let $u(u = 1, \ldots, U)$ be the index of the split. Let $\hat{\delta}_j^{(u)}$ represent the selection indicator estimated from the training data for the uth split. For each region $\mathcal{S}_j$ in the image, we define the stability of selection as the frequency for selecting the location: $U^{-1}\sum_{u=1}^{U}\hat{\delta}_j^{(u)}$.

5.1 Synthetic Nonlinear Classification

We first consider a synthetic two-dimensional nonlinear binary classification example. We generate images $\mathbf{X}_i$ on the $\{1,\dots,30\}\times\{1,\dots,30\}$ grid. The binary response Y_i is defined by

$$Y_i = \begin{cases} 1, & 6\sin\left\{1+\dfrac{3\sum_{j\in\mathcal{J}_1}X_i(\mathcal{S}_j)}{|\mathcal{J}_1|}\right\} - 3\exp\left\{\dfrac{\sum_{j\in\mathcal{J}_2}X_i(\mathcal{S}_j)}{|\mathcal{J}_2|}\right\} - 2 > 0, \\ 0, & \text{otherwise}, \end{cases}$$

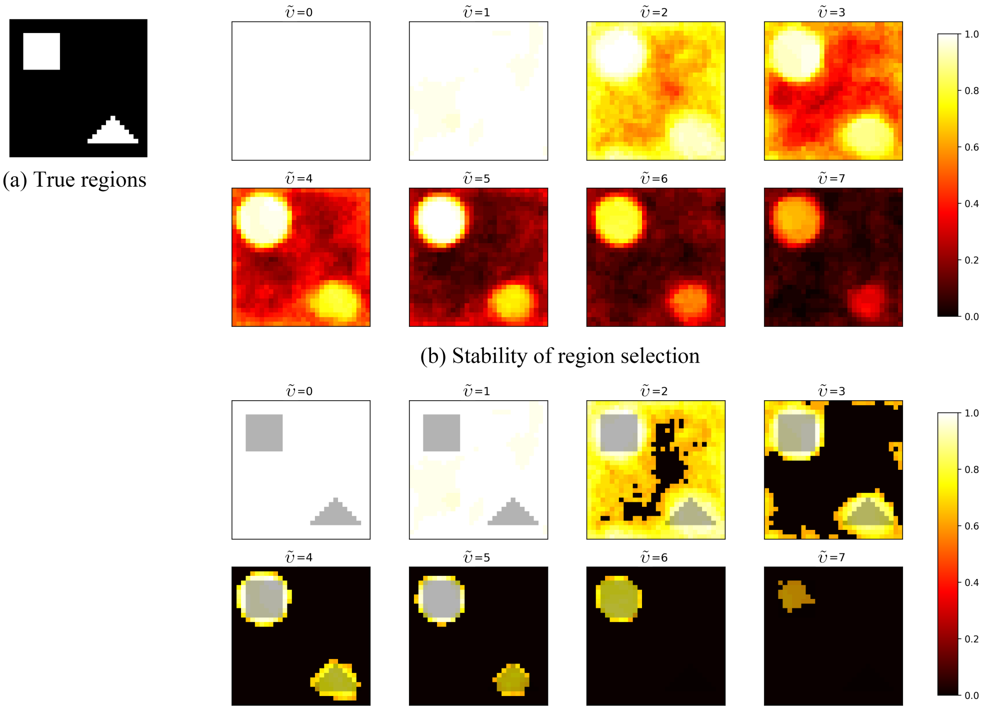
where $\mathcal{J}_1$ and $\mathcal{J}_2$ are the index sets of pixels of square region and triangular region respectively in Figure 2 (a). To make the problem more challenging, we let the covariates $X_i(\mathcal{S}_j)$ share a complex covariance structure, satisfying $X_i(\mathcal{S}_j) = \tilde{X}_i(\mathcal{S}_j)/2 + \epsilon_{ij}I(j\in\mathcal{J}_1\cup\mathcal{J}_2)/2$, where $\tilde{X}_i(\mathcal{S}_j)$ is Gaussian with mean 0 and covariance structure $\mathrm{Cov}\{\tilde{X}_i(\mathcal{S}_j),\tilde{X}_i(\mathcal{S}_l)\} = \exp(-\|s_j - s_l\|_2^2/2)$, ϵ_{ij} are i.i.d. Gaussian variables and s_j is the centroid of the region $\mathcal{S}_j$. The label distribution under this generative is almost balanced. To show that SV-NN is capable of finding true regions given limited samples, we generate 50 such datasets with each dataset consisting of 5000 samples, and set the training set size to 200, with the remaining samples being the test set.

We set $K = 5, \tilde{v} = 5$ in SV-NN and compare its performance with BNN (Liang et al., 2018), STGP (Kang et al., 2018) and CNN. The CNN is constructed to mimic the structure as SV-NN: one 5 × 5 convolution layer with 32 filters, followed by a 2 × 2 max pooling layer and a fully connected layer with 32 hidden units. Table 1 summarizes the error rate of different methods, where we vary the training set size from 20 to 200 while fixing the test set. Among the four methods, SV-NN shows the best performance: it achieves an error rate less than 20% when the training set size is 100 and 200, with the other four above 20%. Note that STGP has an accuracy close to random guess, which is expected given that the underlying data generation model is non-linear. This implies that STGP may only work for the linear case.

Table 1: Binary Classification Error Rate (%) on the Nonlinear Synthetic Dataset under different size of training data.

Training set size	SV-NN	BNN	STGP	CNN
20	**35.4(8.0)**	43.7(10.5)	49.2(4.1)	40.2(9.7)
50	**26.3(7.2)**	39.4(13.2)	48.6(3.7)	31.5(10.1)
100	**19.8(4.7)**	32.2(12.5)	48.3(3.5)	24.6(7.2)
200	**15.5(2.4)**	25.6(9.8)	47.9(3.4)	20.4(4.4)

Next we test the sensitivity of threshold parameter $\tilde{v}$, number of hidden units K and choices of the kernel in the Gaussian process. We vary one of $K, \tilde{v}$ and the kernel, while keeping the other two fixed. The training set size is fixed to be 200. In our experiments on K and $\tilde{v}$, we report results of both 5-fold cross-validation for hyperparameter selection and standard training on the whole training set. The parameters of other kernels (following definitions in Williams and Rasmussen (2006)) are made to match $b = 10$ in the original



(a) True regions

(b) Stability of region selection

(c) Stability of region selection, threshold = 0.6

Figure 2: (a) True regions in the synthetic nonlinear classification example. (b) Stability of region selection of SV-NN with varying threshold $\tilde{v}$ over 50 repetitions on the nonlinear classification example. The expected Bayesian FDR is controlled at 0.01. (c) Selected regions with stability greater than threshold 0.6, overlapped with the true regions.

MSE kernel: $\ell = \sqrt{1/20}$ in squared exponential (SE) kernel, $\ell = \sqrt{1/20}$ and $\nu = 3/2$ in Matérn kernel, and $\ell = \sqrt{1/20}$ in exponential kernel.

Table 2 shows that $\tilde{\upsilon} = 5, K = 5$ with MSE kernel achieves the highest accuracy. Specifically, Table 2(a) implies that accuracy of SV-NN is not sensitive to $\tilde{\upsilon}$ when $\tilde{\upsilon}$ is in a reasonable range (2 to 5 in this synthetic example), but SV-NN may show a poor performance when $\tilde{\upsilon}$ is too large or too small. Stability of region selection under different $\tilde{\upsilon}$s are displayed in Figure 2 (b) and Figure 2(c). We can see that as $\tilde{\upsilon}$ increases, number of selected pixels decreases. All pixels are selected when $\tilde{\upsilon} = 0, 1$ and the lower right triangle region is not selected when $\tilde{\upsilon} = 6, 7$. Regarding the sensitivity with respect to K, Table 2(b) demonstrates that we should use a reasonable number of hidden units in SV-NN. When $K = 1$ (i.e. STGP) and $K = 3$, the neural network fails to capture the non-linear relationship between X and Y. When $K = 7$ and $K = 10$, the accuracy drops a bit given we only have 200 training samples and it is likely for SV-NN to overfit. We observe that choosing a larger K does not hurt the performance too much, while a small K may results in unsatisfactory results. Finally, Table 2(c) reveals that the MSE kernel performs better than SE, Matérn and Exponential kernel.

Table 2: Results of SV-NN on synthetic dataset with different threshold parameter $\tilde{\upsilon}$, number of hidden units K and kernels. Both 5-fold cross validation and standard training on the whole training set are considered for experiments on K and $\tilde{\upsilon}$.

(a) $K = 5$, MSE kernel

υ	Error Rate	
	CV	Test
0	35.4(11.2)	38.5(7.4)
1	19.3(6.4)	19.9(3.2)
2	18.1(6.0)	17.5(3.8)
3	17.3(6.1)	16.7(4.5)
4	16.0(5.7)	15.8(4.2)
5	**15.7(5.3)**	**15.5(2.4)**
6	24.4(12.5)	22.1(11.7)
7	30.8(12.2)	28.4(13.1)

(b) $\tilde{\upsilon} = 5$, MSE kernel

K	Error Rate	
	CV	Test
1	40.8(9.5)	47.9(3.4)
3	26.8(13.0)	29.2(11.3)
5	**15.7(5.3)**	**15.4(2.4)**
7	17.4(6.5)	16.5 (3.0)
10	17.9(8.4)	17.2(7.1)

(c) $K = 5, \tilde{\upsilon} = 5$

Kernel	Error Rate
MSE	**15.4(2.4)**
SE	17.9(3.5)
Matérn	18.6(3.7)
Exponential	19.2(3.8)

5.2 MNIST

The MNIST data contains $60,000$ training images and $10,000$ test images, each being a 28×28 image of handwritten digits. We choose several pairs of digits and use SV-NN as the classifier. To show the performance of our model with limited training samples, we randomly select $n = 10, 20, 30, 40, 50$ samples from each digit in the training set and evaluate classification errors in the test set. For comparison, we also report the classification errors using the whole training set.

In this experiment, we aim to show the advantage of SV-NN when the network structure is shallow and simple, thus we only use $K = 8$ hidden units with one hidden layer. Other parameters are specified as follows: The soft threshold parameter $\tilde{\upsilon} = 2$. Parameters in the modified square exponential kernel are: $a = 0.01, b = 10$, and the polynomial degree of numerical approximation is 20. We run the SGLD algorithm for 5,000 iterations and summarize the posterior inference results using the last 100 samples. For each sample size n, the best step size is chosen from $\{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and each minibatch consists of $\min\{n/2, 128\}$ samples.

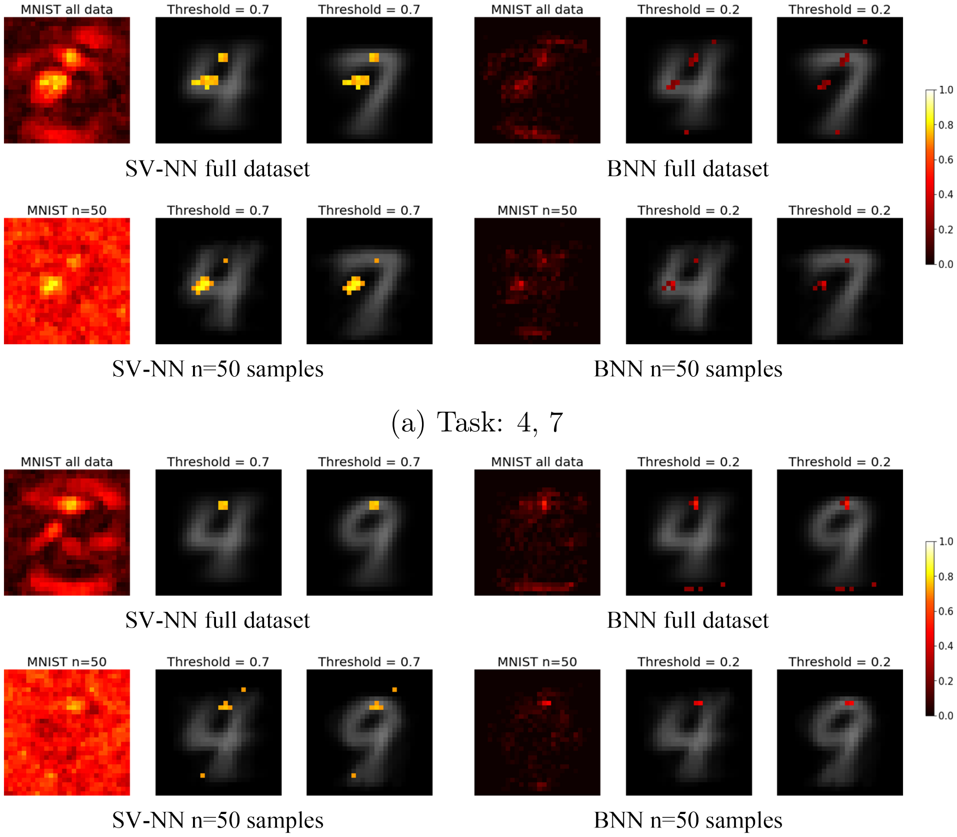
Table 3: Binary Classification Error Rate(%) on MNIST Data

Method	Task	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	All
SV-NN	3, 5	**14.8(3.7)**	**12.5(3.1)**	**10.1(1.8)**	**9.1(1.7)**	**8.7(1.4)**	1.3
	3, 8	**13.3(3.4)**	**11.4(2.7)**	**9.5(1.9)**	**9.0(1.9)**	8.6(2.0)	1.6
	4, 7	**4.6(1.5)**	**4.1(1.3)**	**3.6(1.0)**	**3.4(0.8)**	**3.0(0.6)**	0.7
	4, 9	**14.6(4.1)**	**12.0(2.8)**	**10.1(2.3)**	**9.6(1.6)**	**8.7(1.6)**	1.7
BNN	3, 5	27.1(5.7)	20.8(4.0)	17.9(3.2)	15.6(3.0)	15.0(2.9)	8.8
	3, 8	18.2(4.9)	15.8(4.3)	13.6(3.4)	12.0(2.6)	11.0(2.2)	9.4
	4, 7	18.3(5.6)	13.1(4.1)	10.0(3.0)	8.9(2.5)	8.2(2.5)	4.2
	4, 9	18.8(3.8)	17.4(3.7)	15.9(2.9)	14.9(2.8)	13.7(2.0)	7.6
DNN	3, 5	21.4(4.0)	19.3(5.0)	17.7(6.1)	13.8(6.2)	12.4(5.6)	1.0
	3, 8	21.1(4.8)	17.0(6.0)	14.2(6.7)	10.5(5.9)	8.5(4.8)	1.2
	4, 7	20.8(5.3)	15.2(6.9)	9.8(7.1)	7.7(6.1)	4.1(3.5)	0.8
	4, 9	23.7(2.1)	21.0(4.2)	19.8(5.0)	18.2(5.7)	16.5(5.8)	1.3
CNN	3, 5	18.3(4.0)	15.0(3.4)	12.3(2.9)	11.3(2.8)	10.4(2.3)	**0.9**
	3, 8	14.9(3.9)	11.6(2.4)	10.0(2.3)	9.2(1.6)	**8.3(1.3)**	**0.6**
	4, 7	8.7(3.4)	5.4(1.8)	4.4(1.4)	3.8(1.0)	3.6(0.8)	**0.2**
	4, 9	24.1(5.6)	19.2(4.9)	15.8(4.1)	14.6(3.8)	12.9(3.1)	**0.8**

We compare our SV-NN with BNN (Liang et al., 2018), CNN, and DNN under different sample size settings. The competing models are specified to have a similar architecture as SV-NN. For BNN and DNN, we consider a single hidden layer with 8 hidden units. For CNN, we use only one convolution layer with 8 filters (kernel = 3×3). In Table 3 we show the accuracy of our SV-NN and the competing methods on MNIST.

When we only have limited samples, our model outperforms other neural networks with similar structures, given that it conducts classification and region selection simultaneously. When the whole training set is used, SV-NN also shows comparable performances with its competitors. In fact, MNIST data is not the most appropriate dataset for our model, since the digits are not perfectly centered and the spatial structures appear in a curve-like region instead of a clustered region.

Our method is more interpretable in terms of region selection. Figure 3 shows the stability of region selection over 50 repetitions with SV-NN and BNN using the full dataset and 50 samples. The selection results of SV-NN clearly explain the difference between digit 4 and digit 7. Such selection is very consistent, and has relatively high stability (> 0.7). Specifically, regions selected by SV-NN are those where two digits have distinguishable lines and curves, while the selected regions of BNN are less interpretable and consistent. These results accord with our cognition of the difference between digit 4 and digit 7. Our analysis suggests the distinction between images, and interpretability of SV-NN contributes to the understanding of handwritten data.



(a) Task: 4, 7

(b) Task: 4, 9

Figure 3: Stability of region selection of binary classification over 50 repetitions with SV-NN and BNN using the full dataset (the 1st row) and $n = 50$ samples (the 2nd row), overlapped with the digits. The expected Bayesian FDR is controlled at 0.01.

Table 4: Binary Classification Error Rate(%) on Fashion-MNIST Data

Method	Task	$n = 10$	$n = 20$	$n = 30$	$n = 40$	$n = 50$	All
SV-NN	0, 2	**6.7(1.1)**	6.5(1.1)	6.1(0.8)	5.8(0.7)	5.7(0.7)	3.0
	0, 6	**21.8(1.7)**	21.6(1.4)	**21.0(1.1)**	**20.5(0.9)**	**19.7(0.9)**	13.6
	5, 7	15.7(2.3)	13.3(2.0)	11.0(1.9)	10.0(1.4)	9.4(1.1)	1.9
	7, 9	**8.9(1.3)**	8.4(1.2)	7.9(0.9)	7.5(1.0)	7.1(0.8)	2.5
DNN	0, 2	22.4(6.0)	23.5(5.0)	20.6(7.9)	15.6(9.5)	8.4(7.2)	**1.8**
	0, 6	23.8(3.4)	24.0(3.3)	23.3(4.2)	21.4(5.7)	20.1(6.1)	**7.7**
	5, 7	23.7(3.5)	24.3(3.1)	23.2(4.9)	20.8(6.3)	19.4(6.7)	1.5
	7, 9	20.1(7.4)	21.2(7.2)	17.1(9.1)	10.1(7.5)	**5.0(1.6)**	**2.2**
CNN	0, 2	7.2(1.9)	**6.3(1.0)**	**6.0(1.0)**	**5.7(0.8)**	**5.6(0.7)**	3.1
	0, 6	22.5(2.3)	**21.5(1.7)**	21.0(1.3)	20.7(1.4)	20.2(1.3)	10.6
	5, 7	**15.1(3.2)**	**11.2(2.3)**	**9.0(1.4)**	**7.9(1.2)**	**7.4(1.2)**	**1.1**
	7, 9	9.0(1.9)	**8.1(1.4)**	**7.5(1.0)**	**7.2(0.9)**	7.0(1.1)	2.4

Labels: 0-T-shirt/top, 2-Pullover, 5-Sandal, 6-Shirt, 7-Sneaker, 9-Ankle boot.

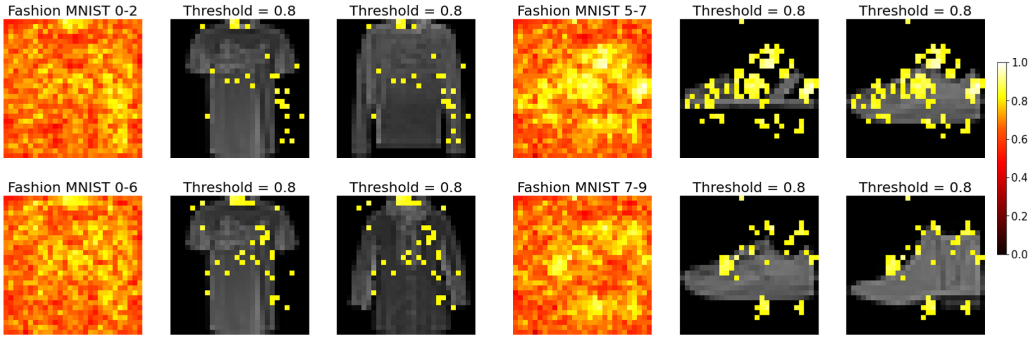


Figure 4: Stability of region selection of binary classification of SV-NN using the full dataset over 50 repetitions, overlapped with the fashion MNIST clothing images. The expected Bayesian FDR is controlled at 0.01.

5.3 Fashion-MNIST

The Fashion-MNIST data (Xiao et al., 2017) is a dataset of Zalando's article images, which contains 60,000 training examples and 10,000 testing examples, each being a 28x28 grayscale image, associated with a label from 10 classes. Similarly, we select several pairs of classes and train SV-NN to classify them. Different from the previous experiment on MNIST, we aim to show the performance of SV-NN when the model architecture is more complex. Here we use $K = 128$ in our SV-NN layer and add another fully connected layer with 64 hidden nodes. Pre-specified parameters are the same as in the previous experiment. For comparison, we also train two-convolutional-layer CNN and two-hidden-layer DNN (128-64). BNN is not computationally tractable with such complexity of the networks.

Table 4 shows the binary classification errors of SV-NN and other models. Although our overall accuracy is lower than that of CNN and DNN training on the whole dataset, again, SV-NN outperforms or is comparable to other models when the training size is relatively small. This implies that our model works well with limited data, which is exactly the case in the neuroimaging field. More importantly, SV-NN achieves region selection simultaneously. Figure 4 shows the stability of image regions selected by SV-NN in multiple binary classification tasks, in which we can find the regions that help distinguish different classes of data are consistently selected with the stability more than 0.8.

5.4 ABCD fMRI data

The Adolescent Brain Cognitive Development (ABCD) study (Casey et al., 2018) is the largest longitudinal study of brain development and child health in the United States. The ABCD study collects longitudinal measurements of brain functions and structures from 9-10 years old across 21 sites in the U.S. One scientific question of interest is to study the associations between cognitive ability, measured by the child's cognitive score (Deary et al., 2007), and working memory brain activity, measured by the task fMRI. In the ABCD study, the working memory task fMRI is collected based on the emotional n-back tasks which engage processes related to memory and emotion regulation (Casey et al., 2018). There are two types of working memory tasks: the 0-back task and 2-back task corresponding to low and high memory load conditions respectively. We prepossess the fMRI data using the ABCD imaging data standard preprocessing pipeline (Casey et al., 2018; Sripada et al., 2020) and construct three contrast maps (2-back versus 0-back, 2-back versus baseline, and 0-back versus baseline) using the statistical parametric mapping (Penny et al., 2011). All the contrast maps are registered into the Montreal Neurological Institute (MNI) 2-mm standard template of dimension $91 \times 109 \times 91$.

Our analysis aims to identify the important imaging biomarkers from the task fMRI contrast maps to make a prediction on the g-factor (Jensen, 1998) which is a typical measure of general intelligence in psychometric investigations of cognitive abilities. We apply the Automated anatomical labeling (AAL) atlas as the brain mask and focus on a total of 185,405 voxels in the contrast map as the potential imaging predictors. In addition, we include p-factor (Caspi et al., 2014), sex, race, and four other covariates as linear predictors. The data we analyze contains 1855 subjects after the imaging data quality control and removing the missing values. For the three contrast maps, we construct three different models and perform the analysis separately.

We compare four different models: the baseline linear model (without brain image as a predictor), SV-NN, and two 3D CNN models. In SV-NN, we set $K = 32$, $b = 100$, and $\tilde{v} = 10$ with other parameters the same as in previous experiments. Two CNN models have similar architectures: Convolution 3d - Max Pooling 3d - Convolution 3d -Max Pooling 3d. For the first CNN (CNN1), two 3d convolutions have 8 filters (kernel $= 3 \times 3$, stride $= 2$) and 16 filters (kernel $= 4 \times 4$, stride $= 2$) respectively, with the flattened output from the second convolution directly connected to the response variable. For the second CNN (CNN2), convolution layers are similar but with more filters (16 and 32), followed by an additional fully connected layer with 32 hidden nodes. The numbers of parameters in SV-NN, CNN1, and CNN2 are 56755, 9778, and 197234, respectively.

Two different ways of data splitting are considered in our experiment. The first is to randomly split the data into 80% for training and 20% for testing. The second is to train the model only using samples from any single site with over 100 subjects and testing the

Table 5: Predictive R-squared and predictive mean squared error (PMSE) of Different Models on ABCD data

Modality	Method	Random Split		Single site	
		R^2	MSE	R^2	MSE
No Image	Linear	0.243(0.032)	0.557(0.035)	0.150(0.055)	0.665(0.163)
2-back vs. 0-back	SV-NN	**0.320(0.014)**	**0.497(0.011)**	**0.195(0.041)**	**0.614(0.075)**
	CNN1	0.307(0.028)	0.524(0.026)	0.186(0.053)	0.629(0.121)
	CNN2	0.311(0.025)	0.518(0.024)	0.182(0.047)	0.640(0.093)
2-back vs. baseline	SV-NN	**0.301(0.020)**	**0.526(0.019)**	**0.184(0.046)**	**0.631(0.101)**
	CNN1	0.289(0.034)	0.533(0.039)	0.177(0.056)	0.647(0.131)
	CNN2	0.288(0.036)	0.531(0.040)	0.179(0.045)	0.649(0.105)
0-back vs. baseline	SV-NN	0.284(0.012)	**0.534(0.013)**	**0.166(0.057)**	**0.653(0.128)**
	CNN1	0.272(0.029)	0.540(0.033)	0.159(0.068)	0.661(0.174)
	CNN2	**0.286(0.027)**	0.535(0.027)	0.164(0.048)	0.655(0.141)

prediction accuracy on the rest of subjects. Table 5 reports the predictive R-squared and mean squared error (MSE) of four models using three different types of contrast maps. In the random-split scenario, SV-NN has the highest R-squared and the lowest MSE in most cases, while CNN2 has comparable results. In the single-site case with limited training data, SV-NN outperforms all other models. Note that the 2-back versus 0-back contrast map is the most informative one among the three contrast maps, and image regions selected by SV-NN in this modality can be used to analyze which brain regions have strong associations with the g-factor. For illustrative purposes, scatter plots depicting the predicted versus actual cognitive scores from a single run are shown in Figure 5.

Similar to the other two data analysis, we run SV-NN 50 times and calculate the stability of region selection with expected Bayesian FDR controlled at 0.01. In Table 6, we summarize the top 10 regions having the largest number of selected voxels with the stability more than 0.6. Their functionality is identified according to AAL regions.

Table 6: Top 10 regions with the largest number of selected voxels. A voxel is selected if its stability is over 0.6. The expected Bayesian FDR of 0.01.

Region Name	# Selected Voxels
Precuneus_L	1471
Precuneus_R	1373
Frontal_Mid_L	1317
Cingulum_Mid_R	1170
Frontal_Sup_L	1062
Cingulum_Mid_L	1056
Frontal_Sup_R	981
Frontal_Mid_R	897
Temporal_Mid_L	851
Lingual_R	828

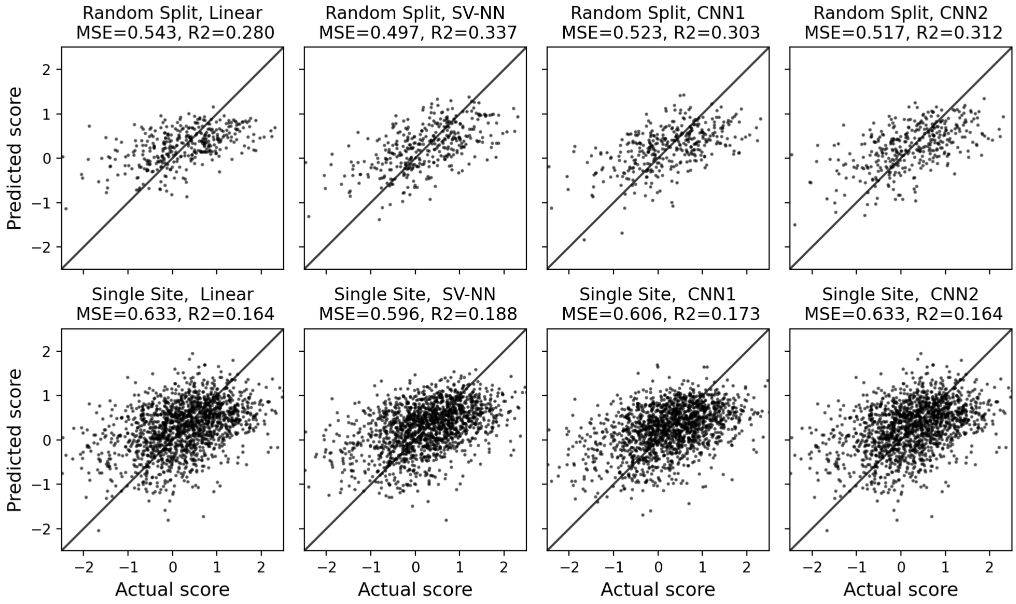
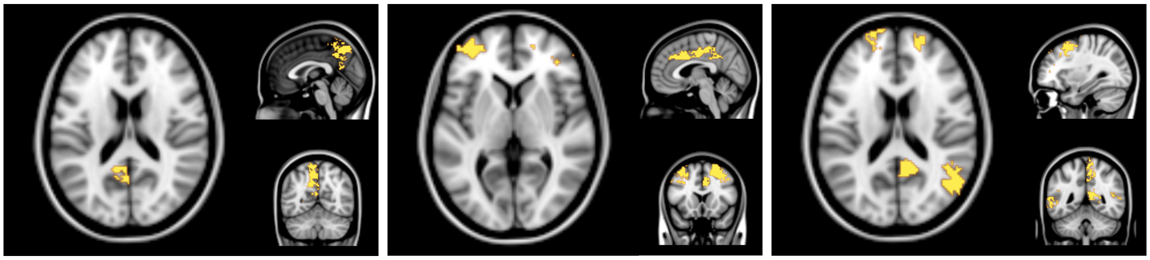


Figure 5: Scatter plots illustrating the relationship between predicted scores and actual scores across four models under the random split and single-site configurations. The models SV-NN, CNN1, and CNN2 employ the 2-back versus 0-back task modality.



(a) Memory Retrieval

(b) Task Control

(c) Default Mode

Figure 6: Voxels selected by SV-NN in the top 10 regions described in Table 6. They are grouped according to their functionality in AAL.

Among the top 10 regions, region Precuneus_L has the largest number of voxels and it is associated with memory retrieval, which matches our expectation since the experiment is on working memory. Most of the other regions are associated with either the task control network(Frontal_Mid_L, Frontal_Sup_R, Frontal_Mid_R) or the default mode network (Precuneus_R, Precuneus_L, Frontal_Sup_L, Frontal_Mid_R, Temporal_Mid_R). In addition, region Ligual_R is related to the visual functional network. Figure 6 presents the top activation regions that are grouped by the major functional networks.

6. Conclusion

In summary, SV-NN is a novel neural network prior for Bayesian scalar-on-image regression. It models the hidden unit specific spatially varying weights that connect the hidden layer to the input imaging predictors with the soft-thresholded Gaussian process (STGP), which can effectively recover the sparse, piecewise smooth and continuous spatial effects of imaging predictors on the response variable of interest. The spatial smoothness and sparsity ensure more direct interpretations of model fitting results of SV-NN compared to those of CNN or DNN. For each hidden unit, SV-NN obtains a sparse spatially varying weight estimate that clearly indicates the important image regions that may contribute to the prediction. The overlap between the selected regions across different hidden units may reflect the interactions between different types of imaging features. In contrast, general DNN methods may impose the sparsity on the weight parameters but ignore the spatial dependence between image pixels or voxels. Although CNN improves DNN by constructing the spatially varying features with the kernel convolution approach, it mainly focuses on local spatial features as the kernel size is typically small and it does not directly select the spatially varying features.

We perform rigorous theoretical analysis for a single layer SV-NN. We establish the model identifiability conditions which define the space of the imaging effect function, the space of underlying signal intensity and the desired properties of the activation functions. Those results imply practical guidelines for model specifications, image data preprocessing and choice of the activation functions. For example, Condition 3 suggests that the commonly used sigmoid and ReLU functions can be adopted for feature selection in SV-NN. From the Bayesian nonparametric inference perspective, we also establish the posterior consistency of the imaging effect function and selection consistency of the feature regions of input images, providing theoretical guarantees for response variable prediction and feature selection in scalar-on-image regression. For theoretical convenience, we assume that the dispersion parameter of the exponential family, which is typically related to the variance of the distribution, is known in our model. This assumption simplifies the derivation of posterior properties and facilitates analytical tractability. However, the proposed framework can be extended to accommodate an unknown dispersion parameter, following similar strategies as in Choi (2005).

For posterior computation, we adopt an equivalent model representation based on the basis expansion approximation to Gaussian processes. Although this approach is general for any covariance kernel with numerical approximations, we particularly discuss implementations of the modified squared exponential kernel with a closed form of the eigen functions and eigen values, which can be utilized to efficiently implement the MCMC algorithm for large-scale image data. From our experience, the performance of SV-NN is not very sensitive to the choices of kernel parameters as long as they are specified within a reasonable range. Depending on the applications, one may consider different kennels such as the Matérn kernel and the gamma-exponential kernel, which may be more suitable for some special spatial smoothness patterns. For the MCMC methods, we choose a gradient-based approach, i.e., SGLD, which is scalable to the analysis of the data with a large sample size. The gradient approximation for the soft-threshold operator works well, as it is in the same fashion for ReLU function in many deep neural network applications. For the motivating applications in this paper, SGLD performs well if the step size is tuned in a reasonable range. One may consider other stochastic gradient MCMC methods including SGHMC to speed up the convergence of MCMC according to the needs of applications. In the Bayesian framework, the Posterior Inclusion Probability (PIP) serves as a measure of uncertainty regarding the inclusion of specific image regions in the model. The reliability of PIPs hinges on accurate model specification, appropriate prior selection, and the behavior of the posterior distribution. Our theoretical analysis demonstrates the proposed model's flexibility, the soundness of the SV-NN prior, and the posterior consistency in region selection. Consequently, PIPs can be viewed as subjective probabilistic assessments within the context of the assumed model. Unlike classical deep neural networks, which do not inherently offer model-level uncertainty quantification, our approach benefits from the Bayesian framework's strengths. By employing PIPs alongside a two-stage thresholding procedure, we effectively identify significant regions. Stability analysis in experiments further supports the effectiveness of PIPs as indicators of uncertainty and highlights the robustness of our inference procedure.

We apply SV-NN to analyze three motivating datasets: the MNIST, Fashion MNIST, and ABCD data. SV-NN shows comparable prediction accuracy compared to classic DNN and CNN models when the training data has a large sample size. When the training sample size is small, a single layer SV-NN with limited number of hidden units often has a better prediction performance compared to CNN or DNN with more complicated structures and involving more parameters. In addition, SV-NN can select regions that contain key features of images that are strongly associated with the response variable. The selected regions based on a small training sample size can be very close to the selected regions based on a large training sample size. For example, in the analysis of the MNIST data, SV-NN uses 100 training samples (50 for each digit) to identify two spatial contiguous regions that capture the essential differences between digits 4 and 7. The selected regions based on all 12,107 training samples (5842 for digit 4, 6265 for digit 7) remain similar. In the analysis of ABCD data, SV-NN achieves an overall better prediction accuracy compared to the two CNNs and selects scientifically meaningful brain regions that can contribute to predicting the general cognitive ability.

Acknowledgements

Ben Wu's research was partially supported by National Natural Science Foundation of China (NSFC, 12201628) and the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001).

Appendix A. Proofs

A.1 Proof of Lemmas

A.1.1 Proof of Lemma 1

Proof We follow Theorem 3.1 of Mhaskar (2020) to show the approximation error bound between $f^*(.)$ and $\psi(.;\boldsymbol{\theta}^*)$. Specifically, we show that the image kernel $\mathcal{H}(.,.)$ satisfies with the following smoothness conditions: a) $\|\mathcal{H}(\mathbf{X},.)-\mathcal{H}(\mathbf{X}',.)\|_\infty \le C\|\mathbf{X}-\mathbf{X}'\|_1$; and b) $\sup_{\mathbf{X}\in\mathbb{X}}\|\mathcal{H}(\mathbf{X},.)\|_{\mathbb{B}_p,1}<\infty$, where

$$\|\mathcal{H}(\mathbf{X},.)\|_{\mathbb{B}_p,1} = \sup_{\boldsymbol{\beta}\in\mathbb{B}_p}\ \sup_{\mathbf{b}\in O_c(\boldsymbol{\beta}),0<\delta\le c} \frac{\inf_{P\in\Pi_1}\|\mathcal{H}(\mathbf{X},.)-P\|_{O_\delta(\mathbf{b}),\infty}}{\delta},$$

Π_1 is a space of polynomials of degree 1, $c>0$ is a constant, and $\|.\|_{O_\delta(\mathbf{b}),\infty}$ is the supremum norm on the ball $O_\delta(\mathbf{b})=\{\boldsymbol{\beta}\in\mathbb{B}_p : \|\boldsymbol{\beta}-\mathbf{b}\|_1<\delta\}$. For any $\boldsymbol{\beta}\in\mathbb{B}_p$, by Condition 5, we have

$$\|\boldsymbol{\beta}\|_2^2=\sum_{j=1}^p\{\beta(\mathcal{S}_j)\}^2=\sum_{j=1}^p\left\{\int_{\mathcal{S}_j}\phi(s)m(ds)\right\}^2\le\sum_{j=1}^p p\{m(\mathcal{S}_j)\}^2\le L^{-2}.$$

By Condition 3, there exists a constant C such that

$$\begin{aligned}|\mathcal{H}(\mathbf{X},\boldsymbol{\beta})-\mathcal{H}(\mathbf{X}',\boldsymbol{\beta})| &= \left|h\left\{\sum_{j=1}^p\beta(\mathcal{S}_j)X(\mathcal{S}_j)\right\}-h\left\{\sum_{j=1}^p\beta(\mathcal{S}_j)X'(\mathcal{S}_j)\right\}\right|\\ &\le M_0\left|\sum_{j=1}^p\beta(\mathcal{S}_j)\left\{X(\mathcal{S}_j)-X'(\mathcal{S}_j)\right\}\right|\\ &\le M_0L^{-1}\|\mathbf{X}-\mathbf{X}'\|_2\le C\|\mathbf{X}-\mathbf{X}'\|_1,\end{aligned}$$

which indicates that condition a) holds. Next, let

$$P_{\mathbf{b}}(\boldsymbol{\beta})=\mathcal{H}(\mathbf{X},\mathbf{b})+\sum_{j=1}^p X(\mathcal{S}_j)\left\{\beta(\mathcal{S}_j)-b(\mathcal{S}_j)\right\}.$$

By Conditions 3 and 4, for any $\boldsymbol{\beta}\in O_\delta(\mathbf{b})$,

$$\begin{aligned}\frac{|\mathcal{H}(\mathbf{X},\boldsymbol{\beta})-P_{\mathbf{b}}(\boldsymbol{\beta})|}{\delta} \le& \frac{M_0+1}{\delta}\left|\sum_{j=1}^p\{\beta(\mathcal{S}_j)-b(\mathcal{S}_j)\}X(\mathcal{S}_j)\right|\\ \le& \frac{M_0+1}{\delta}\|\mathbf{X}\|_2\|\boldsymbol{\beta}-\mathbf{b}\|_2\le(M_0+1)M_1.\end{aligned}$$

Therefore,

$$\sup_{\mathbf{X}\in\mathbb{X}}\|\mathcal{H}(\mathbf{X},.)\|_{\mathbb{B}_p,1}=\sup_{\mathbf{X}\in\mathbb{X}}\sup_{\boldsymbol{\beta}\in\mathbb{B}_p}\ \sup_{\mathbf{b}\in O_c(\boldsymbol{\beta}),0<\delta\le c}\frac{\inf_{P\in\Pi_1}\|\mathcal{H}(\mathbf{X},.)-P\|_{O_\delta(\mathbf{b}),\infty}}{\delta}<\infty,$$

which indicates that condition b) holds. According to Theorem 3.1 of Mhaskar (2020), we have

$$\|f - \psi(.;\boldsymbol{\theta})\|_\infty \leq CM_b \left(\frac{\log K}{K}\right)^{1/2} K^{-1/r}.$$

■

A.1.2 Proof of Lemma 2

Proof Noting that

$$\pi_n(\mathbf{D}_{n,i}; f) = \pi(Y_i|\mathbf{X}_i, \mathbf{W}_i, \boldsymbol{\alpha}, f)\pi(\mathbf{X}_i)\pi(\mathbf{W}_i),$$

then

$$\begin{aligned}&\Lambda_n(\mathbf{D}_{n,i}; f^*, f)\\ =&\log \pi_n(\mathbf{D}_{n,i}; f^*) - \log \pi_n(\mathbf{D}_{n,i}; f)\\ =&Y_i\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\} + b\left\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i)\right\} - b\left\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\}.\end{aligned}$$

By the tower property of conditional expectation and the mean value theorem, we have

$$\begin{aligned}&E_{f^*}\left\{\Lambda_n(\mathbf{D}_{n,i}; f^*, f)\right\}\\ =&E_{f^*}\left(\left[b'\left\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\} - b'(\tilde{\gamma}_i)\right]\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\}\right),\end{aligned}$$

where $\min\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i), \mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\} < \tilde{\gamma}_i < \max\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i), \mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\}$. By Condition 3, $b'(.)$ is continuously differentiable. Then there exists a constant C such that,

$$\left[b'\left\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\} - b'(\tilde{\gamma}_i)\right]\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\} \leq C\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\}^2.$$

Let $\tilde{\mathcal{F}} = \{f : \|f - f^*\|_\infty < \xi\}$, for any $f \in \tilde{\mathcal{F}}$, we have

$$\frac{1}{n}\sum_{i=1}^n K_{n,i}(f^*, f) \leq \frac{1}{n}\sum_{i=1}^n E_{f^*}\left[C\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\}^2\right] \leq C\xi^2.$$

By Theorem 1, for any $\xi < \varepsilon^{1/2}C^{-1}$,

$$\Pr\left(f \in \tilde{\mathcal{F}}, \frac{1}{n}\sum_{i=1}^n K_{n,i}(f^*, f) < \varepsilon\right) > 0.$$

In addition, by Conditions 3 and 4,

$$\begin{aligned}&E_{f^*}\left[Y_i\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\} + b\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f(\mathbf{X}_i)\} - b\{\mathbf{W}_i^\top\boldsymbol{\alpha} + f^*(\mathbf{X}_i)\}\right]^2\\ \leq&2(M_1 + C^2)E_{f^*}\left\{f^*(\mathbf{X}_i) - f(\mathbf{X}_i)\right\}^2 \leq 2(M_1 + C^2)\xi^2.\end{aligned}$$

Then,

$$\begin{aligned}V_{n,i}(f^*, f) =&\mathrm{Var}_{f^*}\left\{\Lambda_n(\mathbf{D}_{n,i}; f^*, f)\right\}\\ \leq&2(M_1 + C^2)\xi^2.\end{aligned},$$

and

$$\frac{1}{n^2}\sum_{i=1}^{n} V_{n,i}(f^*, f) \to 0, \quad n \to \infty, \quad f \in \tilde{\mathcal{F}}.$$

■

A.1.3 Proof of Lemma 3

Proof Denote $\mathcal{N}_n$ as the vector space of $\boldsymbol{\psi} := (\psi(\mathbf{X}_1;\boldsymbol{\theta}),\ldots,\psi(\mathbf{X}_n;\boldsymbol{\theta}))^\top$ with $\boldsymbol{\theta} = \{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K,\boldsymbol{\zeta}\}$, $\boldsymbol{\beta}_k \in \mathbb{B}_p$, and $|\zeta_k| < M_b$, $k = 1,\ldots,K$. Denote $\mathcal{L}_n$ as the vector space of $(L(\mathbf{X}_1,\boldsymbol{\beta}),\ldots,L(\mathbf{X}_n,\boldsymbol{\beta}))^\top$ with $L(\mathbf{X},\boldsymbol{\beta}) = \sum_{j=1}^{p}\beta(\mathcal{S}_j)X(\mathcal{S}_j)$ and $\boldsymbol{\beta} \in \mathbb{B}_p$. Then, for any $\boldsymbol{\psi}^* \in \mathcal{N}_n$, we have

$$\begin{aligned}
&\{\boldsymbol{\psi} \in \mathcal{N}_n : \|\boldsymbol{\psi} - \boldsymbol{\psi}^*\|_\infty < \varepsilon\}\\
\supset &\bigcap_{k=1}^{K}\left\{\max_i |\zeta_k\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k) - \zeta_k^*\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)| < \frac{\varepsilon}{K}\right\}\\
\supset &\bigcap_{k=1}^{K}\left\{\max_i |\zeta_k\left\{\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k) - \mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)\right\}| < \frac{\varepsilon}{2K}\right\} \cap \left\{\max_i |(\zeta_k - \zeta_k^*)\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)| < \frac{\varepsilon}{2K}\right\}\\
\supset &\bigcap_{k=1}^{K}\left\{\max_i |\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k) - \mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)| < \frac{\varepsilon}{2KM_b}\right\} \cap \left\{|\zeta_k - \zeta_k^*| < \frac{\varepsilon}{2K\max_i |\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)|}\right\}\\
\supset &\bigcap_{k=1}^{K}\left\{\max_i \left|\sum_{j=1}^{p}\{\beta_k(\mathcal{S}_j)X_i(\mathcal{S}_j) - \beta_k^*(\mathcal{S}_j)X_i(\mathcal{S}_j)\}\right| < \frac{\varepsilon}{2KM_0M_b}\right\} \cap \left\{|\zeta_k - \zeta_k^*| < \frac{\varepsilon}{2K\max_i |\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)|}\right\}.
\end{aligned}$$

By Condition 3, h is Lipschitz continuous, then

$$\begin{aligned}
|\mathcal{H}(\mathbf{X},\boldsymbol{\beta})| &\le |\mathcal{H}(\mathbf{X},\boldsymbol{\beta}) - \mathcal{H}(\mathbf{0},\boldsymbol{\beta})| + |\mathcal{H}(\mathbf{0},\boldsymbol{\beta})|\\
&\le M_0\|\mathbf{X}\|_2\|\boldsymbol{\beta}\|_2 + |h(0)|.
\end{aligned}$$

By Theorem 4 of Zhang (2002), if $\|\mathbf{X}_i\|_2 \le b$, $i = 1,\ldots,n$ and $\|\boldsymbol{\beta}_k\|_2 \le a$, $k = 1,\ldots,K$, there exist a constant C such that

$$\begin{aligned}
&\log N(\varepsilon,\mathcal{N}_n,\|.\|_\infty)\\
\le &K\log N\left(\frac{\varepsilon}{2KM_0M_b},\mathcal{L}_n,\|.\|_\infty\right) + K\log\left\{\frac{2M_bK\max_i |\mathcal{H}(\mathbf{X}_i,\boldsymbol{\beta}_k^*)|}{\varepsilon}\right\}\\
\le &\frac{4Ca^2b^2K^2M_0^2M_b^2}{\varepsilon^2}\log\left(\frac{2abKM_0M_bn}{\varepsilon}\right) + K\log\left\{\frac{2M_bK(abM_0 + |h(0)|)}{\varepsilon}\right\}.
\end{aligned}$$

Let $\mathcal{M}_n := \{(f(\mathbf{X}_1),\ldots,f(\mathbf{X}_n))^\top : f \in \mathcal{F}_p\}$ and $\mathbf{f} := (f(\mathbf{X}_1),\ldots,f(\mathbf{X}_n))^\top$, for any $f \in \mathcal{F}_p$, $\mathbf{v} := (v_1,\ldots,v_n) \in \mathcal{M}_n$, $\|\mathbf{f} - \mathbf{v}\|_\infty \le \|\mathbf{f} - \boldsymbol{\psi}\|_\infty + \|\boldsymbol{\psi} - \mathbf{v}\|_\infty$. By Lemma 1, for a large K we have

$$\|\mathbf{f} - \boldsymbol{\psi}\|_\infty \le CM_b\left(\frac{\log K}{K}\right)^{1/2}K^{-1/r} < \frac{\varepsilon}{2}.$$

Then $\|\boldsymbol{\psi}-\mathbf{v}\|_\infty < \varepsilon/2$ implies $\|\mathbf{f}-\mathbf{v}\|_\infty < \varepsilon$, which means $N(\varepsilon, \mathcal{M}_n, \|.\|_\infty) \leq N(\varepsilon/2, \mathcal{N}_n, \|.\|_\infty)$.

By Condition 7, there exist constant C_1 such that

$$\begin{aligned}\log N\left(\varepsilon, \mathcal{F}_p, \|.\|_\infty\right) \leq & \log \sup_{\mathbf{X}_1,\ldots,\mathbf{X}_n\in\mathbb{X}} N\left(\frac{\varepsilon}{2}, \mathcal{N}_n, \|.\|_\infty\right) \\ \leq & \frac{16CL^{-2}K^2M_0^2M_1^2M_b^2}{\varepsilon^2}\log\left(\frac{4L^{-1}KM_0M_1M_bn}{\varepsilon}\right) \\ & + K\log\left\{\frac{2M_bK(M_0M_1L^{-1}+|h(0)|)}{\varepsilon}\right\} \\ \leq & C_1 n^{1-\nu/2}\varepsilon^{-2}.\end{aligned}$$

■

A.1.4 Proof of Lemma 4

Proof Notice that for any $f \in \mathcal{F}$, we have

$$f(\mathbf{X}) = \int_{\mathbb{B}} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta})\tau(d\boldsymbol{\beta}) = \int_{\mathbb{B}_p} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta})\tau(d\boldsymbol{\beta}) + \int_{\mathbb{B}\setminus\mathbb{B}_p} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta})\tau(d\boldsymbol{\beta}).$$

Then

$$\mathcal{F}_p^C = \left\{ f\in\mathcal{F} : \left| \int_{\mathbb{B}\setminus\mathbb{B}_p} \mathcal{H}(\mathbf{X}, \boldsymbol{\beta})\tau(d\boldsymbol{\beta}) \right| > 0 \right\}.$$

For any $f \sim \mathcal{SV}\text{-}\mathcal{NN}(K, \sigma_\zeta^2, \upsilon, \kappa)$, we can write

$$f(\mathbf{X}) = \psi(\mathbf{X};\boldsymbol{\theta}) = \frac{1}{K}\sum_{k\in\mathcal{K}_1} \zeta_k \mathcal{H}(\mathbf{X}, \boldsymbol{\beta}_k) + \frac{1}{K}\sum_{k\in\mathcal{K}_2} \zeta_k \mathcal{H}(\mathbf{X}, \boldsymbol{\beta}_k),$$

where $\boldsymbol{\beta}_k = (\beta_k(\mathcal{S}_1), \ldots, \beta_k(\mathcal{S}_p))^\top$, $\beta_k(\mathcal{S}_j) = \int_{\mathcal{S}_j} \phi_k(s)m(ds)$, $j = 1, \ldots, p$, and $\mathcal{K}_1$ and $\mathcal{K}_2$ are two index sets which are defined as $\mathcal{K}_1 := \{k : \|\phi_k\|_\infty \leq p^{1/2}\}$ and $\mathcal{K}_2 := \{k : \|\phi_k\|_\infty > p^{1/2}\}$. Then, it's obvious that $\mathcal{K}_2 = \emptyset$ implies $f \in \mathcal{F}_p$. Then for any $f \in \mathcal{F}_p^C$, $\mathcal{K}_2$ is not empty, which implies

$$\Pr\left(f\in\mathcal{F}_p^C\right) \leq \sum_{k=1}^K \Pr\left(\|\phi_k\|_\infty > p^{1/2}\right) \leq C_0K\exp\left(-C_1p\right).$$

The last inequality is obtained by Theorem 5 of Ghosal and Roy (2006) and Lemma 4 of Kang et al. (2018). By Condition 7, we then have

$$\Pr\left(f\in\mathcal{F}_p^C\right) \leq C_0\exp\left(-C_1n\right).$$

■

A.1.5 Proof of Lemma 5

Proof Following Choi (2005), we define

$$d(f_1, f_2) := \inf\left\{\varepsilon : Q(\{\mathbf{X} : |f_1(\mathbf{X}) - f_2(\mathbf{X})| > \varepsilon\}) < \varepsilon\right\},$$

where Q is the marginal probability measure of $\mathbf{X}$. Since $\|f - f^*\|_1 > \varepsilon$ and by Condition 4, π_x is bounded away from zero, then there exists ε_1 such that

$$\int_{\mathbb{X}} |f(\mathbf{X}) - f^*(\mathbf{X})| Q(d\mathbf{X}) = \int_{\mathbb{X}} |f(\mathbf{X}) - f^*(\mathbf{X})| \pi_x(\mathbf{X}) m(d\mathbf{X}) > \varepsilon_1,$$

which implies $d(f, f^*) > \varepsilon_2$ for some $\varepsilon_2 > 0$ considering f and f^* are both bounded functions (by Conditions 3 and 4). Let

$$\mathcal{D} = \left\{\mathbf{x} : |f^*(\mathbf{x}) - f(\mathbf{x})| > \varepsilon_2\right\},$$

then $\Pr(\mathcal{D}) > \varepsilon_2$. By the mean value theorem, there exists γ_i such that $\min\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i), \mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\} < \gamma_i < \max\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i), \mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\}$, and

$$\left| g^{-1}\left\{\mathbf{W}^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\} - g^{-1}\left\{\mathbf{W}^\top \boldsymbol{\alpha} + f(\mathbf{X}_i)\right\} \right| = |b''(\gamma_i)| \left| f^*(\mathbf{X}_i) - f(\mathbf{X}_i) \right|.$$

Now, by Condition 3, since b'' is continuous and $|b''(\gamma_i)| > 0$ for any γ_i, it is easy to see that for any constant $M_\gamma > 0$, $\inf_{|\gamma_i| \le M_\gamma} |b''(\gamma_i)| > 0$. Noting that for any $f \in \mathcal{F}_p$,

$$|f(\mathbf{X}_i)| \le \int_{\mathbb{B}_p} |\mathcal{H}(\mathbf{X}_i, \boldsymbol{\beta})| |\tau|(d\boldsymbol{\beta}) \le M_0 \|\mathbf{X}_i\|_2 \int_{\mathbb{B}_p} \|\boldsymbol{\beta}\|_2 |\tau|(d\boldsymbol{\beta}) \le M_0 M_1 M_b L^{-1}.$$

Let $\tilde{M} := \min_i \inf_{|\gamma_i| \le M_\gamma} |b''(\gamma_i)|$ and $M_\gamma = M_0 M_1 M_b L^{-1} + \max_i |\mathbf{W}_i^\top \boldsymbol{\alpha}|$, then we have

$$\begin{aligned}
\Pr\left(\mathcal{A}_n^C\right) =& \Pr\left[\mathbf{X}_i : \sum_{i=1}^n \left| g^{-1}\left\{\mathbf{W}^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\} - g^{-1}\left\{\mathbf{W}^\top \boldsymbol{\alpha} + f(\mathbf{X}_i)\right\} \right| < n\epsilon\right] \\
\le& \Pr\left\{\varepsilon_2 \sum_{i=1}^n b''(\gamma_i) I(\mathbf{X}_i \in \mathcal{D}) < n\epsilon\right\} \\
\le& \Pr\left\{\varepsilon_2 \sum_{i=1}^n I(\mathbf{X}_i \in \mathcal{D}) < \frac{n\epsilon}{\tilde{M}}\right\} + \sum_{i=1}^n \Pr\left(|\gamma_i| > M_\gamma\right) \\
\le& \Pr\left\{\varepsilon_2 \sum_{i=1}^n I(\mathbf{X}_i \in \mathcal{D}) < \frac{n\epsilon}{\tilde{M}}\right\} \\
&+ \sum_{i=1}^n \Pr\left\{|f(\mathbf{X}_i)| > M_0 M_1 M_b L^{-1}\right\} + \sum_{i=1}^n \Pr\left\{|f^*(\mathbf{X}_i)| > M_0 M_1 M_b L^{-1}\right\} \\
=& \Pr\left\{\varepsilon_2 \sum_{i=1}^n I(\mathbf{X}_i \in \mathcal{D}) < \frac{n\epsilon}{\tilde{M}}\right\}.
\end{aligned}$$

Let $Z = n - \sum_{i=1}^{n} I\left\{\mathbf{X}_i \in \mathcal{D}\right\}$. Then Z follows a binomial distribution with parameters n and $1 - \Pr(\mathcal{D})$, and is stochastically dominated by a binomial distribution with parameters n and $1 - \varepsilon_2$, denoted as $\tilde{Z}$. Then, let

$$\begin{aligned} t &= \log\left[\{\varepsilon_2^2 M_b - \varepsilon_2\epsilon\}/\{\epsilon(1-\varepsilon_2)\}\right], \\ C_0 &= \log\left\{\frac{\epsilon}{\varepsilon_2^2 M_b}\right\} + t\left\{1 - \frac{\epsilon}{\varepsilon_2 M_b}\right\}, \end{aligned}$$

and by Markov's inequality, we have

$$\begin{aligned} \Pr\left(\mathcal{A}_n^C\right) \leq & \Pr\left\{\mathbf{X}_i : \varepsilon_2 \sum_{i=1}^{n} I(\mathbf{X}_i \in \mathcal{D}) < \frac{n\epsilon}{M_b}\right\} \\ \leq & \Pr\left[\tilde{Z} > n\left\{1 - \frac{\epsilon}{\varepsilon_2 M_b}\right\}\right] \\ = & \Pr\left(\exp\left(t\tilde{Z}\right) > \exp\left[tn\left\{1 - \frac{\epsilon}{\varepsilon_2 M_b}\right\}\right]\right) \\ \leq & \{\varepsilon_2 + (1-\varepsilon_2)\exp(t)\}^n \exp\left[-tn\left\{1 - \frac{\epsilon}{\varepsilon_2 M_b}\right\}\right] \\ \leq & \exp\left(-C_0 n\right). \end{aligned}$$

■

A.1.6 Proof of Lemma 6

Proof For the hypothesis testing problem

$$H_0 : f = f^*, \quad H_1 : f = f^{**},$$

we construct the test statistic

$$\Psi_n\left(f^*, f^{**}\right) = I\left(\sum_{i=1}^{n} \delta_i \left[Y_i - g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\}\right] > \frac{1}{2} r_0 n\right),$$

where $\delta_i = 1$, if $g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i)\right\} - g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\} > 0$, $\delta_i = -1$ otherwise. By Hoeffding's inequality,

$$\begin{aligned} & E_{f^*}\left\{\Psi_n\left(f^*, f^{**}\right)\right\} \\ = & \Pr_{f^*}\left(\sum_{i=1}^{n} \delta_i \left[Y_i - g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\}\right] > \frac{1}{2} r_0 n\right) \leq \exp(-r_0^2 n/2). \end{aligned}$$

By Lemma 5, there exists $N_0 > 0$ such that for all $n > N_0$, with probability one,

$$\sum_{i=1}^{n} \left| g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i)\right\} - g^{-1}\left\{\mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i)\right\}\right| \geq r_0 n.$$

Then for any f such that $\|f - f^{**}\|_\infty \leq r_0/(4M_0)$,

$$\begin{aligned}
&E_f \left\{1 - \Psi_n\left(f^*, f^{**}\right)\right\} \\
=& \Pr_f \left(\sum_{i=1}^n \delta_i \left[Y_i - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i) \right\} \right] \leq \frac{1}{2} r_0 n \right) \\
=& \Pr_f \left(\sum_{i=1}^n \delta_i \left[Y_i - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} \right] \right. \\
&+ \delta_i \left[g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i) \right\} \right] \\
&\left. + \delta_i \left[g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i) \right\} - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i) \right\} \right] \leq \frac{1}{2} r_0 n \right) \\
\leq& \Pr_f \left(\sum_{i=1}^n \delta_i \left[Y_i - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} \right] \right. \\
&- \left| g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i) \right\} \right| \\
&\left. + \left| g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^{**}(\mathbf{X}_i) \right\} - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f^*(\mathbf{X}_i) \right\} \right| \leq \frac{1}{2} r_0 n \right) \\
\leq& \Pr_f \left(\sum_{i=1}^n \delta_i \left[Y_i - g^{-1}\left\{ \mathbf{W}_i^\top \boldsymbol{\alpha} + f(\mathbf{X}_i) \right\} \right] \leq -\frac{1}{4} r_0 n \right) \leq \exp(-r_0^2 n/8).
\end{aligned}$$

Then we can construct uniformly consistent tests Ψ_n. Specifically, let $t = \min\{r_0/(4M_0), \varepsilon/2\}$, and $f^1, \ldots, f^{N_t} \in \mathcal{F}_p$, such that for any $f \in \mathcal{F}_p$ there exist at least one l such that $\|f^l - f\|_\infty < t$. Define

$$\Psi_n = \max_{1 \leq l \leq N_t} \Psi_n(f^*, f^l).$$

If $\|f - f^*\|_1 > \varepsilon$, then $\|f^l - f^*\|_1 > \varepsilon/2$ for f^l which satisfies with $\|f - f^l\|_\infty < t \leq \varepsilon/2$. Then by Lemma 5,

$$E_{f^*}\{\Psi_n(f^*, f^l)\} \leq \exp(-r_0^2 n/2),$$
$$E_f\{1 - \Psi_n(f^*, f^l)\} \leq \exp(-r_0^2 n/8).$$

Therefore, we can find a large enough N such that for any $n > N$,

$$\begin{aligned}
E_{f^*}(\Psi_n) &\leq \sum_{l=1}^{N_t} E_{f^*}\{\Psi_n(f^*, f^l)\} \leq \exp(\log N_t - r_0^2 n/2) \\
&\leq \exp\left(16 M_0^2 C_1 n^{1-\nu/2} r_0^{-2} - r_0^2 n/2\right) \leq \exp\left(-r_0^2 n/4\right),
\end{aligned}$$

and

$$E_f(1 - \Psi_n) = E_f \left\{ 1 - \max_{1 \leq l \leq N_t} \Psi_n(f^*, f^l) \right\} \leq \exp(-r_0^2 n/8).$$

■

A.2 Proof of Theorems and Corollary

A.2.1 Proof of Theorem 1

Proof Notice that

$$
\begin{aligned}
&\{\|\psi(.;\boldsymbol{\theta}) - f^*\|_\infty < \varepsilon\} \\
&\supseteq \left\{\|\psi(.,\boldsymbol{\theta}) - \psi(.,\boldsymbol{\theta}^*)\|_\infty < \frac{\varepsilon}{2}\right\} \bigcap \left\{\|\psi(.,\boldsymbol{\theta}^*) - f^*\|_\infty < \frac{\varepsilon}{2}\right\}.
\end{aligned}
$$

By Lemma 1, for any $f^* \in \mathcal{F}$ and $\varepsilon > 0$, we can find large enough p and K such that $f^* \in \mathcal{F}_p$ and $\|\psi(.;\boldsymbol{\theta}^*) - f^*\|_\infty < \varepsilon/2$. Then, we only need to evaluate the event $\{\|\psi(.;\boldsymbol{\theta}) - \psi(.;\boldsymbol{\theta}^*)\|_\infty < \varepsilon/2\}$. Theorem 1 in Kang et al. (2018) showed that under the smoothness Condition 6, soft-thresholded Gaussian process has large support over the space $\mathcal{I}$, that is, for any function $\phi_k^* \in \mathcal{I}$ and $\varepsilon > 0$, the soft-thresholded Gaussian process prior $\phi_k \sim \mathrm{STGP}(\upsilon_0, \kappa)$ satisfies $\Pr(\|\phi_k - \phi_k^*\|_\infty < \varepsilon) > 0$. In addition, since $\zeta_k \sim \mathcal{N}(0, \sigma_\zeta^2)$, we have $\Pr(\|\zeta_k\|_\infty < \varepsilon) > 0$ and $\Pr(\|\zeta_k - \zeta_k^*\|_\infty < \varepsilon) > 0$. Then,

$$
\begin{aligned}
&\Pr\left(\|\psi(.;\boldsymbol{\theta}) - \psi(.;\boldsymbol{\theta}^*)\|_\infty < \varepsilon\right) \\
=&\Pr\left(\|K^{-1}\sum_{k=1}^{K}\{\zeta_k\mathcal{H}(.,\boldsymbol{\beta}_k) - \zeta_k^*\mathcal{H}(.,\boldsymbol{\beta}_k^*)\}\|_\infty < \varepsilon\right) \\
\geq&\prod_{k=1}^{K}\Pr\left(K^{-1}\|\zeta_k\mathcal{H}(.,\boldsymbol{\beta}_k) - \zeta_k^*\mathcal{H}(.,\boldsymbol{\beta}_k^*)\|_\infty < \frac{\varepsilon}{K}\right) \\
\geq&\prod_{k=1}^{K}\Pr\left(\|\zeta_k\left\{\mathcal{H}(.,\boldsymbol{\beta}_k) - \mathcal{H}(.,\boldsymbol{\beta}_k^*)\right\}\|_\infty < \frac{\varepsilon}{2}\right)\Pr\left(\|\left(\zeta_k - \zeta_k^*\right)\mathcal{H}(.,\boldsymbol{\beta}_k^*)\|_\infty < \frac{\varepsilon}{2}\right).
\end{aligned}
$$

By Condition 3, $h(.)$ is Lipschitz continuous, then

$$
\begin{aligned}
&\left\{\|\mathcal{H}(.,\boldsymbol{\beta}_k) - \mathcal{H}(.,\boldsymbol{\beta}_k^*)\|_\infty < \frac{\varepsilon}{2}\right\} \\
&\supseteq \left\{\sup_{\mathbf{X}\in\mathbb{X}} M_0\left|\sum_{j=1}^{p}\{\beta_k(\mathcal{S}) - \beta_k^*(\mathcal{S}_j)\}X(\mathcal{S}_j)\right| < \frac{\varepsilon}{2}\right\} \\
&\supseteq \left\{\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^*\|_2 \sup_{\mathbf{X}\in\mathbb{X}}\|\mathbf{X}\|_2 < \frac{\varepsilon}{2M_0}\right\}.
\end{aligned}
$$

By Condition 4, $\mathbf{X}$ has a bounded norm, and by Condition 5,

$$
\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_k^*\|_2^2 = \sum_{j=1}^{p}\left\{\int_{\mathcal{S}_j}\phi_k(s) - \phi_k^*(s)m(ds)\right\}^2 \leq \sum_{j=1}^{p}\|\phi_k - \phi_k^*\|_\infty^2\left\{m(\mathcal{S}_j)\right\}^2 \leq L^{-2}p^{-1}\|\phi_k - \phi_k^*\|_\infty^2,
$$

Then, we have

$$
\Pr\left(\|\mathcal{H}(.,\boldsymbol{\beta}_k) - \mathcal{H}(.,\boldsymbol{\beta}_k^*)\|_\infty < \frac{\varepsilon}{2}\right) > 0.
$$

Meanwhile, $\|\mathcal{H}(.,\boldsymbol{\beta}_k^*)\|_\infty \leq \sup_{\mathbf{X}\in\mathbb{X}} M_0\|\boldsymbol{\beta}_k^*\|_2\|\mathbf{X}\|_2$ is bounded. Therefore,

$$
\Pr\left(\|\psi(.;\boldsymbol{\theta}) - f^*\|_\infty < \varepsilon\right) > 0.
$$

■

A.2.2 Proof of Theorem 2

Proof Theorem 2 can be shown by follwoing Theorem A·1 of Choudhuri et al. (2004). Lemma 2 already shows the condition of prior positivity of neighborhoods. By Lemmas 3 - 6, we can verify the existence of tests:

$$\begin{aligned} E_{f^*}(\Phi_n) &\to 0, \\ \sup_{f\in\mathcal{U}_\varepsilon^C\cap\mathcal{F}_p} E_f(1-\Phi_n) &\le C_0\exp(-C_1 n), \\ \Pr(\mathcal{F}_p^C) &\le C_2\exp(-C_3 n), \end{aligned}$$

where $\mathcal{U}_\varepsilon^C=\{f\in\mathcal{F}:\|f-f^*\|_1>\varepsilon\}$.

■

A.2.3 Proof of Theorem 3

Proof There exists a constant $C>0$ such that

$$\begin{aligned} \Delta_j|f^*| =& \sup_{0<\delta<c}\int_{\mathbb{X}_\delta}|f^*(\mathbf{X})-f^*(\mathbf{X}+\delta\mathbf{e}_j)|\,m(d\mathbf{X}) \\ \le& \Delta_j|f|+\|f^*-f\|_1+\sup_{0<\delta<c}\int_{\mathbb{X}_\delta}|f(\mathbf{X}+\delta\mathbf{e}_j)-f^*(\mathbf{X}+\delta\mathbf{e}_j)|\,m(d\mathbf{X}) \\ \le& \Delta_j|f|+(1+C)\left\|f^*-f\right\|_1. \end{aligned}$$

Similarly we can show $\Delta_j|f|\le\Delta_j|f^*|+(1+C)\|f^*-f\|_1$. Therefore,

$$\sup_j|(\Delta_j|f^*|-\Delta_j|f|)|\le(1+C)\|f^*-f\|_1.$$

By Theorem 2, as $n,p,K\to\infty$, we have

$$\Pr\left(\sup_j|(\Delta_j|f^*|-\Delta_j|f|)|<\varepsilon|\mathbf{D}_n\right)\to 1.$$

■

A.2.4 Proof of Corollary 4

Proof For $f(.)=\psi(.;\boldsymbol{\theta})\sim\mathcal{SV}\text{-}\mathcal{NN}(K,\sigma_\zeta^2,\upsilon,\kappa)$, by Condition 3, we have

$$\begin{aligned} |f(\mathbf{X})-f(\mathbf{X}+\mathbf{e}_j\delta)| &\le K^{-1}\sum_{k=1}^K|\zeta_k|\,|\mathcal{H}(\mathbf{X},\boldsymbol{\beta}_k)-\mathcal{H}(\mathbf{X}+\mathbf{e}_j\delta,\boldsymbol{\beta}_k)| \\ &\le K^{-1}M_0\sum_{k=1}^K|\zeta_k|\,|\beta_k(\mathcal{S}_j)\delta|\,, \end{aligned}$$

which implies

$$\left\{j : \max_{k=1,\ldots,K} |\beta_k(\mathcal{S}_j)| = 0\right\} \subseteq \{j : \Delta_j |f| = 0\}\,.$$

By Theorem 3, we know as $n, p, K \to \infty$, for any $\varepsilon > 0$, with probability one,

$$\{j : \Delta_j |f| = 0\} \subseteq \{j : \Delta_j |f^*| < \varepsilon\}\,.$$

Therefore, for any $\varepsilon > 0$, as $n, p, K \to \infty$,

$$\Pr\left(\mathcal{I}_0 \subseteq \mathcal{I}^*_\varepsilon \mid \mathbf{D}_n\right) \to 1.$$

■

References

Jose M Alvarez and Mathieu Salzmann. Learning the number of neurons in deep networks. In *Advances in Neural Information Processing Systems*, pages 2270–2278, 2016.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *The Journal of Machine Learning Research*, 18(1):629–681, 2017.

Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

BJ Casey, Tariq Cannonier, May I Conley, Alexandra O Cohen, Deanna M Barch, Mary M Heitzeg, Mary E Soules, Theresa Teslovich, Danielle V Dellarco, Hugh Garavan, et al. The adolescent brain cognitive development (abcd) study: imaging acquisition across 21 sites. *Developmental Cognitive Neuroscience*, 32:43–54, 2018.

Avshalom Caspi, Renate M Houts, Daniel W Belsky, Sidra J Goldman-Mellor, HonaLee Harrington, Salomon Israel, Madeline H Meier, Sandhya Ramrakha, Idan Shalev, Richie Poulton, et al. The p factor: one general psychopathology factor in the structure of psychiatric disorders? *Clinical Psychological Science*, 2(2):119–137, 2014.

Tianqi Chen, Emily Fox, and Carlos Guestrin. Stochastic gradient hamiltonian monte carlo. In *International Conference on Machine Learning*, pages 1683–1691, 2014.

Yao Chen, Qingyi Gao, Faming Liang, and Xiao Wang. Nonlinear variable selection via deep neural networks. *Journal of Computational and Graphical Statistics*, 30(2):484–492, 2021.

Taeryon Choi. *Posterior consistency in nonparametric regression problems under Gaussian process priors*. PhD thesis, Carnegie Mellon University, 2005.

Nidhan Choudhuri, Subhashis Ghosal, and Anindya Roy. Bayesian estimation of the spectral density of a time series. *Journal of the American Statistical Association*, 99(468):1050–1059, 2004.

Ian J Deary, Steve Strand, Pauline Smith, and Cres Fernandes. Intelligence and educational achievement. *Intelligence*, 35(1):13–21, 2007.

Jean Feng and Noah Simon. Sparse-input neural networks for high-dimensional nonparametric regression and classification. *arXiv preprint arXiv:1711.07592*, 2017.

Subhashis Ghosal and Anindya Roy. Posterior consistency of gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5):2413–2429, 2006.

Daniel Gianola, Hayrettin Okut, Kent A Weigel, and Guilherme JM Rosa. Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. *BMC Genetics*, 12:1–14, 2011.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 315–323, 2011.

Jeff Goldsmith, Lei Huang, and Ciprian M Crainiceanu. Smooth scalar-on-image regression via spatial bayesian variable selection. *Journal of Computational and Graphical Statistics*, 23(1):46–64, 2014.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18(7):1527–1554, 2006.

Arthur Robert Jensen. *The g factor: The science of mental ability*, volume 648. Praeger Westport, CT, 1998.

Jian Kang, Brian J Reich, and Ana-Maria Staicu. Scalar-on-image regression via the soft-thresholded gaussian process. *Biometrika*, 105(1):165–184, 2018.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Ismael Lemhadri, Feng Ruan, Louis Abraham, and Robert Tibshirani. Lassonet: A neural network with feature sparsity. *Journal of Machine Learning Research*, 22(127):1–29, 2021.

Chunyuan Li, Changyou Chen, David Carlson, and Lawrence Carin. Preconditioned stochastic gradient langevin dynamics for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016a.

Fan Li, Tingting Zhang, Quanli Wang, Marlen Z Gonzalez, Erin L Maresh, James A Coan, et al. Spatial bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics*, 9(2):687–713, 2015.

Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5): 322–336, 2016b.

Faming Liang, Qizhai Li, and Lei Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972, 2018.

Baoyuan Liu, Min Wang, Hassan Foroosh, Marshall Tappen, and Marianna Pensky. Sparse convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 806–814, 2015.

Chao Ma, Lei Wu, et al. The barron space and the flow-induced function spaces for neural network models. *Constructive Approximation*, 55(1):369–406, 2022.

Yan Meng and Pingbing Ming. A new function space from barron class and application to neural network approximation. *Commun. Comput. Phys.*, 32(5):1361–1400, 2022.

Hrushikesh N Mhaskar. Dimension independent bounds for general shallow networks. *Neural Networks*, 123:142–152, 2020.

Jeffrey S Morris, Philip J Brown, Richard C Herrick, Keith A Baggerly, and Kevin R Coombes. Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–489, 2008.

Hayrettin Okut, Daniel Gianola, Guilherme JM Rosa, and Kent A Weigel. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. *Genetics Research*, 93(3):189–201, 2011.

William D Penny, Karl J Friston, John T Ashburner, Stefan J Kiebel, and Thomas E Nichols. *Statistical parametric mapping: the analysis of functional brain images*. Elsevier, 2011.

Armenak Petrosyan, Anton Dereventsov, and Clayton G Webster. Neural network integral representations with the relu activation function. In *Mathematical and Scientific Machine Learning*, pages 128–143. PMLR, 2020.

Nicholas G Polson and Veronika Ročková. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Hippolyt Ritter, Aleksandar Botev, and David Barber. A scalable laplace approximation for neural networks. In *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, volume 6. International Conference on Representation Learning, 2018.

Simone Scardapane, Danilo Comminiello, Amir Hussain, and Aurelio Uncini. Group sparse regularization for deep neural networks. *Neurocomputing*, 241:81–89, 2017.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Michael Smith and Ludwig Fahrmeir. Spatial bayesian variable selection with application to functional magnetic resonance imaging. *Journal of the American Statistical Association*, 102(478):417–431, 2007.

Chandra Sripada, Saige Rutherford, Mike Angstadt, Wesley K Thompson, Monica Luciana, Alexander Weigard, Luke H Hyde, and Mary Heitzeg. Prediction of neurocognition in youth from resting state fmri. *Molecular Psychiatry*, 25(12):3413–3421, 2020.

Yan Sun and Faming Liang. A kernel-expanded stochastic neural network. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):547–578, 2022.

Yan Sun, Wenjun Xiong, and Faming Liang. Sparse deep learning: A new framework immune to local traps and miscalibration. *Advances in Neural Information Processing Systems*, 34: 22301–22312, 2021.

Yan Sun, Qifan Song, and Faming Liang. Consistent sparse deep learning: Theory and computation. *Journal of the American Statistical Association*, 117(540):1981–1995, 2022a.

Yan Sun, Qifan Song, and Faming Liang. Learning sparse deep neural networks with a spike-and-slab prior. *Statistics & Probability Letters*, 180:109246, 2022b.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688, 2011.

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

Stephan Wojtowytsch et al. Representation formulas and pointwise properties for barron functions. *Calculus of Variations and Partial Differential Equations*, 61(2):1–37, 2022.

Ben Wu, Ying Guo, and Jian Kang. Bayesian spatial blind source separation via the thresholded gaussian process. *Journal of the American Statistical Association*, 119(545): 422–433, 2024.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Tong Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2(Mar):527–550, 2002.

Mingyuan Zhou. Parsimonious bayesian deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.