Bayesian Spatially Varying Weight Neural Networks with the Soft-Thresholded Gaussian Process Prior

Jian Kang

Department of Biostatistics
University of Michigan, Ann Arbor

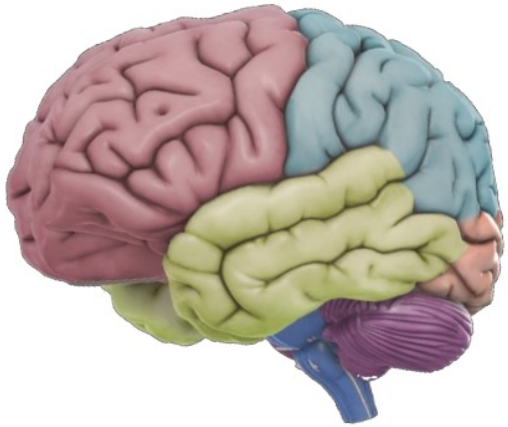
Neuroimaging

EEG



PET/CT





MEG

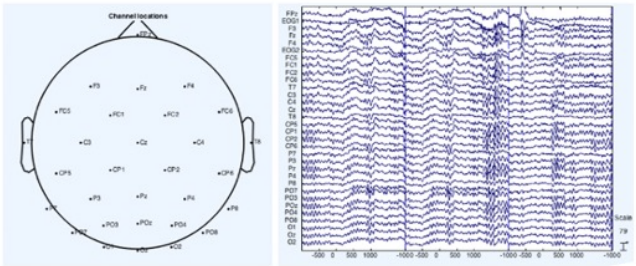


MRI

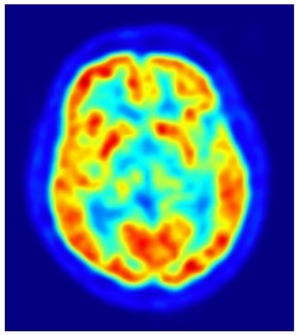


Neuroimaging

EEG



PET

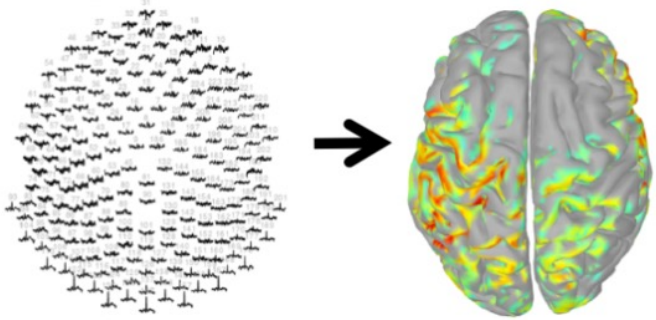


CT

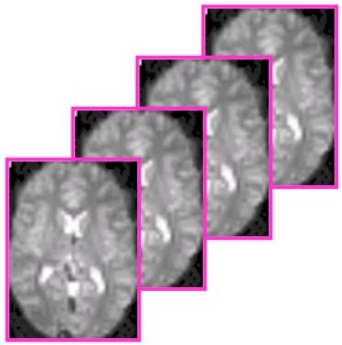




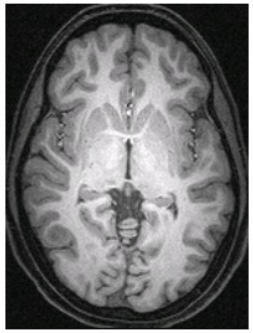
MEG



fMRI

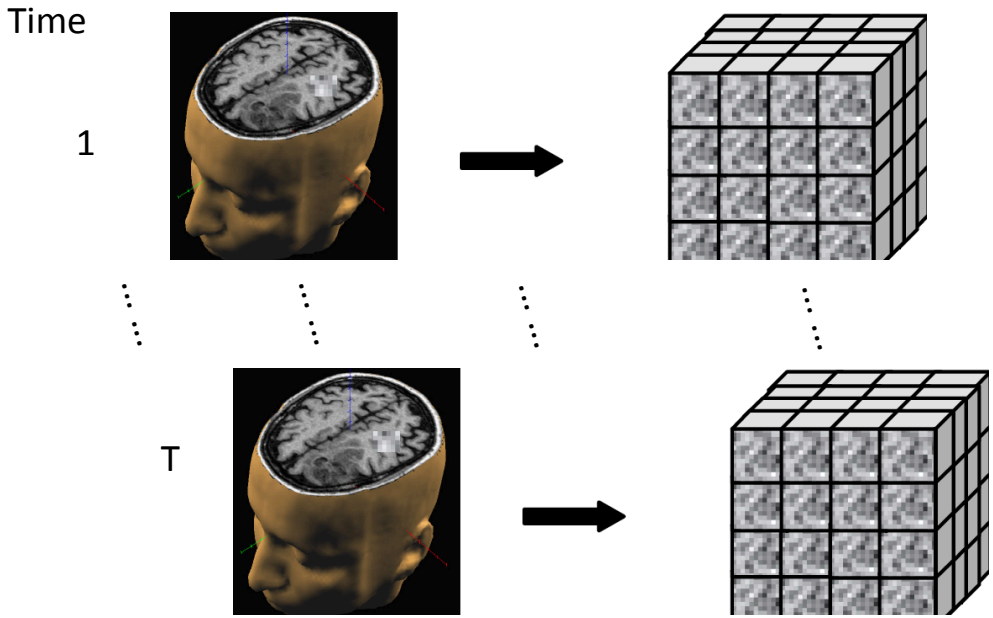


MRI



Data Characteristics

- **3D/4D Data (Spatial-Temporal Data)**



- Massive dat sets: 1,000,000 pixels/voxels in images.
- Important features: contiguous regions, sharp edges and jumps
- Complex spatial correlations (neighbors, long-range between ROIs)
- Temporal correlations

Objectives

Prediction/classification studies: use medical images to

- classify a subject's group membership (e.g. disease status)
- predict clinical response or behavior (e.g. treatment response)
- predict other types of medical images

Challenging problem

- High dimensionality: the number of voxels level imaging predictors is often larger than the number of subjects.
- Spatial/Temporal Dependence: predictors may exhibit complex correlation structure.
- Sparsity: most regions of the image may have no effect on the response.

Challenging problem

- High dimensionality: the number of voxels level imaging predictors is often larger than the number of subjects.
- Spatial/Temporal Dependence: predictors may exhibit complex correlation structure.
- Sparsity: most regions of the image may have no effect on the response.

Solution: Scalar-on-image regression (with feature selection)

Existing Methods: Scalar-on-image Regression

- Frequentist methods:
  - Fussed Lasso[1]
  - Wavelet method[2]
  - Total Variation Norm[3]
- Bayesian methods:
  - Ising Prior[4]
  - Soft-Thresholding Gaussian Process (STGP)[5]
- Limitations: Linear model or Generalized linear model
  - Insufficient to capture the complex association
  - lack of fit; inaccurate region selection and prediction

---

[1] Tibshirani, R. *et al.* Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 91–108 (2005).

[2] Reiss, P. T. *et al.* Wavelet-domain regression and predictive inference in psychiatric neuroimaging. *Annals of Applied Statistics* **9,** 1076 (2015).

[3] Wang, X. *et al.* Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association* **112,** 1156–1168 (2017).

[4] Li, F. *et al.* Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. *The Annals of Applied Statistics* **9,** 687–713 (2015).

[5] Kang, J. *et al.* Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105,** 165–184 (2018).

Deep Learning Methods

- Deep neural networks (DNN)
  - Successful in many computer vision and medical imaging applications[6];
  - Overcome the curse of dimensionality for the nonparametric regression;
  - May achieve high prediction accuracy in complex nonlinear models with **sufficiently large training sample size**;
- Sparse regularizations and feature selection for DNN
  - Control the sparsity of the weights in the first hidden layer with the L_1 penalty[7] or the L_0 penalty[8].
- Convolutional neural network (CNN)[9]
  - A huge breakthrough for image analysis; convolutional layers
  - Difficult to interpret fitting results
  - lack of unified rule to determine the network architecture

---

[6] LeCun, Y. *et al.* Deep learning. *Nature* **521,** 436 (2015).

[7] Lemhadri, I. *et al.* LassoNet: A Neural Network with Feature Sparsity. *Journal of Machine Learning Research* **22,** 1–29 (2021).

[8] Chen, Y. *et al.* Nonlinear Variable Selection via Deep Neural Networks. *Journal of Computational and Graphical Statistics,* 1–9 (2020).

[9] Gu, J. *et al.* Recent advances in convolutional neural networks. *Pattern Recognition* **77,** 354–377 (2018).

Bayesian Neural Networks

- Prior specifications on the weight parameters
  - Gaussian priors[10,11];
  - Spike-and-slab priors[12,13,14].
- Prior specifications on the neural network architecture
  - Gamma process prior to shrink the width and a layer-wise greedy-learning strategy to shrink the depth of the network[15]
- Lack of methods: explicitly account for the spatial dependence among the imaging predictors

---

[10] MacKay, D. J. A practical Bayesian framework for backpropagation networks. *Neural computation* **4,** 448–472 (1992).

[11] Neal, R. M. *Bayesian learning for neural networks.* (Springer Science & Business Media, 2012).

[12] Polson, N. G. & Ročková, V. Posterior concentration for sparse deep learning. *Advances in Neural Information Processing Systems* **31** (2018).

[13] Liang, F. *et al.* Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association* **113,** 955–972 (2018).

[14] Sun, Y. *et al.* Learning sparse deep neural networks with a spike-and-slab prior. *Statistics & Probability Letters,* 109246 (2021).

[15] Zhou, M. Parsimonious Bayesian deep networks. *Advances in neural information processing systems* (2018).

A New Approach: BNN-STGP

- Model: Bayesian neural network (BNN) with STGP priors on
  - Sparse, continuous and piecewise smooth spatially varying weights.
  - Model interpretability: select important image features (regions).
  - Prediction: high accuracy with small training sample size.
- Theoretical analysis: BNN-STGP with a single hidden layer
  - Model identifiability; Large Prior Support; Posterior consistency; Selection consistency
- Numerical Experiments: benchmark digital image datasets
- Neuroimaging Applications: predict IQ from the task fMRI in the ABCD study

Spatially varying weight neural networks

Data:

- Y_i: Scalar outcome, for $i = 1, \ldots, n$;
- $\mathbf{X}_i = \{X_i(s_1), \ldots, X_i(s_p)\}^\top$: image predictors with $s_j \in \mathcal{B}$;
- $\mathbf{W}_i = (W_{i,1}, \ldots, W_{i,q})^\top$: non-image scalar predictors

Model:

- The conditional density belongs to an exponential family, i.e.,

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

- $b(\cdot)$ and $A(\cdot)$: known functions such that
- $g\{E(Y_i \mid \mathbf{X}_i, \mathbf{W}_i)\} = \mathbf{W}_i^\top \boldsymbol{\alpha} + \psi(\boldsymbol{\theta}, \mathbf{X}_i)$ with $g^{-1}(\cdot) = b'(\cdot)$.
- $\boldsymbol{\alpha}$: linear effects of non-image scalar predictors.
- $\psi(\boldsymbol{\theta}, \mathbf{X}_i)$: constructed by a neural network.

Single layer spatially varying weight neural network

One layer neural network consisting of K hidden units:

$$\psi(\boldsymbol{\theta}, \mathbf{X}_i) = \sum_{k=1}^{K} \zeta_k Z_{k,i},$$

$$Z_{k,i} = h\left\{ \eta_k + p^{-1/2} \sum_{j=1}^{p} \beta_k(s_j) X_i(s_j) \right\}.$$

For the kth hidden unit,

- $Z_{k,i}$: non-linear transformation of the ith image predictor.
- ζ_k: linear effect of $Z_{k,i}$ on the scalar outcome.
- $h(\cdot)$: non-linear activation function, e.g., ReLU: $\max\{0, x\}$.
- $\beta_k(s_j)$: spatially varying weights of the image feature at location s_j.
- η_k: bias parameter.

Single layer spatially varying weight neural network

One layer neural network consisting of K hidden units:

$$\psi(\boldsymbol{\theta}, \mathbf{X}_i) = \sum_{k=1}^{K} \zeta_k Z_{k,i},$$

$$Z_{k,i} = h\left\{ \eta_k + p^{-1/2} \sum_{j=1}^{p} \beta_k(s_j) X_i(s_j) \right\}.$$

For the kth hidden unit,

- $Z_{k,i}$: non-linear transformation of the ith image predictor.
- ζ_k: linear effect of $Z_{k,i}$ on the scalar outcome.
- $h(\cdot)$: non-linear activation function, e.g., ReLU: $\max\{0, x\}$.
- $\beta_k(s_j)$: spatially varying weights of the image feature at location s_j.
- η_k: bias parameter.

Why the normalizing scalar $p^{-1/2}$?

Single layer spatially varying weight neural network

One layer neural network consisting of K hidden units:

$$\psi(\boldsymbol{\theta}, \mathbf{X}_i) = \sum_{k=1}^{K} \zeta_k Z_{k,i},$$

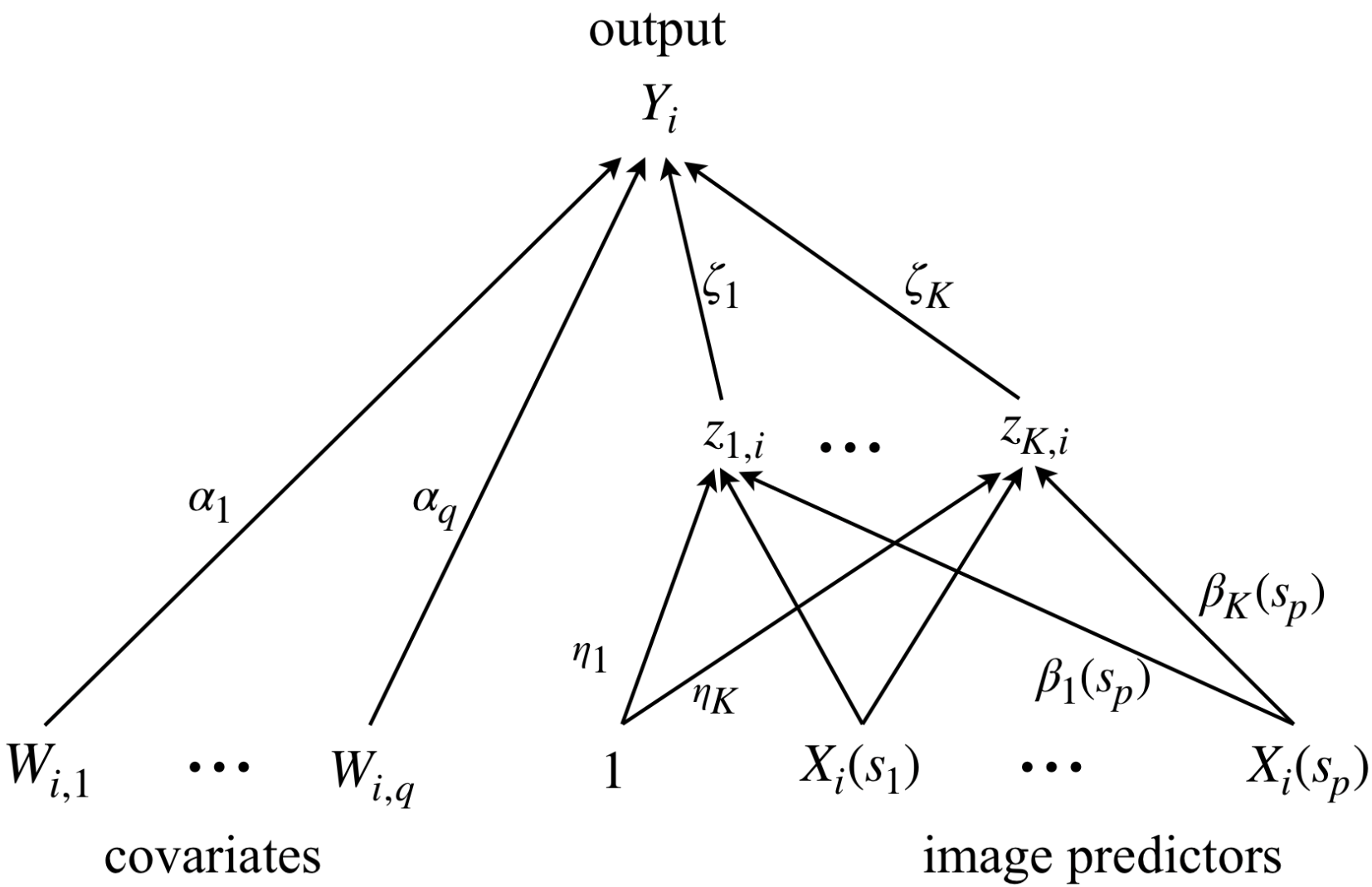
$$Z_{k,i} = h\left\{ \eta_k + p^{-1/2} \sum_{j=1}^{p} \beta_k(s_j) X_i(s_j) \right\}.$$

For the kth hidden unit,

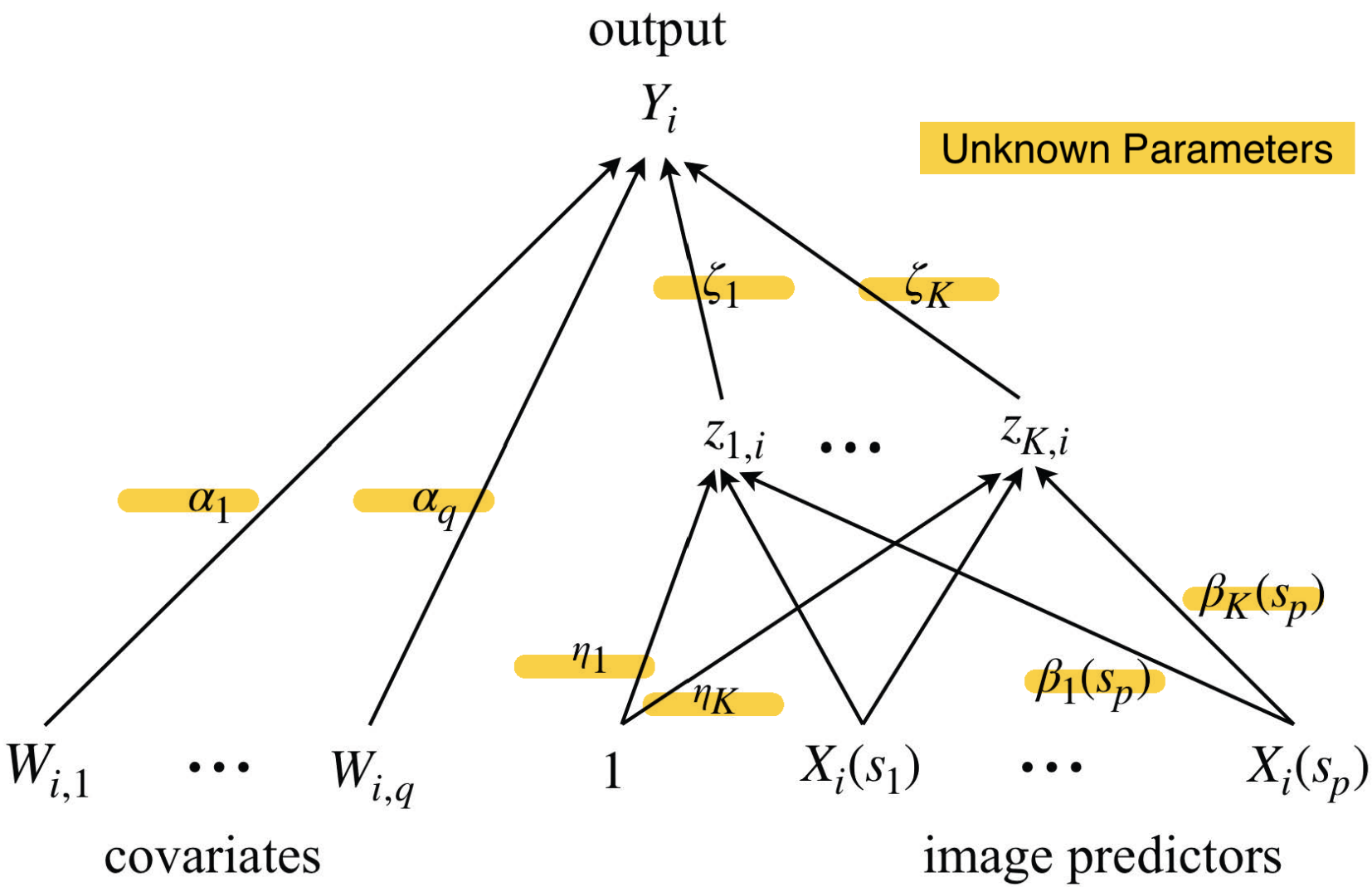
- $Z_{k,i}$: non-linear transformation of the ith image predictor.
- ζ_k: linear effect of $Z_{k,i}$ on the scalar outcome.
- $h(\cdot)$: non-linear activation function, e.g., ReLU: $\max\{0, x\}$.
- $\beta_k(s_j)$: spatially varying weights of the image feature at location s_j.
- η_k: bias parameter.

The total effect of the image predictors reflects the total amount of the intensity in the brain signals, which should not increase to infinity as the image resolution increases

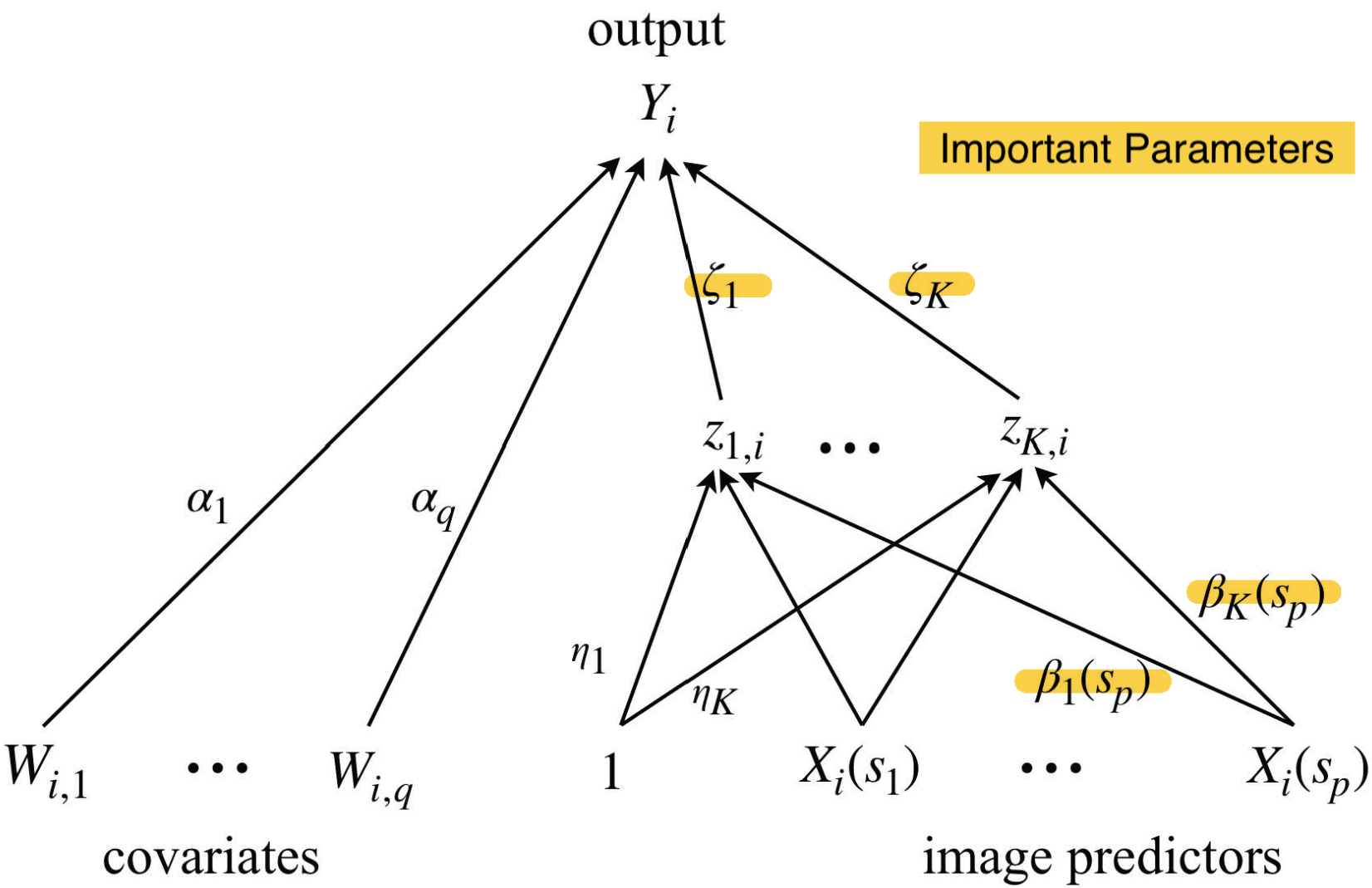
Overview of Spatially varying weight neural networks



Overview of Spatially varying weight neural networks



Overview of Spatially varying weight neural networks



Prior Specifications

- Spatially varying weights:

$$\beta_k(\cdot) \sim \mathcal{STGP}\{\upsilon, \kappa(\cdot,\cdot)\}$$

  with positive prior probability on those events:
  - Sparse: many locations $\mathbf{s}_j$'s have $\beta_k(\mathbf{s}_j) = 0$
  - Spatial: sites with nonzero coefficients cluster spatially
  - Smooth: the coefficients vary smoothly in clusters of nonzero coefficients.

- Non-spatial regression coefficients:

$$\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}_q, \sigma_\alpha^2 \mathbf{I}_q), \quad \boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}_K, \sigma_\zeta^2 \mathbf{I}_K), \quad \boldsymbol{\eta} \sim \mathcal{N}(\mathbf{0}_K, \sigma_\eta^2 \mathbf{I}_K), \quad (1)$$

- Variance parameters:

$\sigma_b^2, \sigma_\alpha^2, \sigma_\zeta^2$ and σ_η^2 follow inverse gamma priors.

Soft Thresholded Gaussian Processes (STGP)

- This STGP[16] prior for spatially varying weights

$$\beta_k(\cdot) \sim \mathcal{STGP}\{\upsilon, \kappa(\cdot,\cdot)\}$$

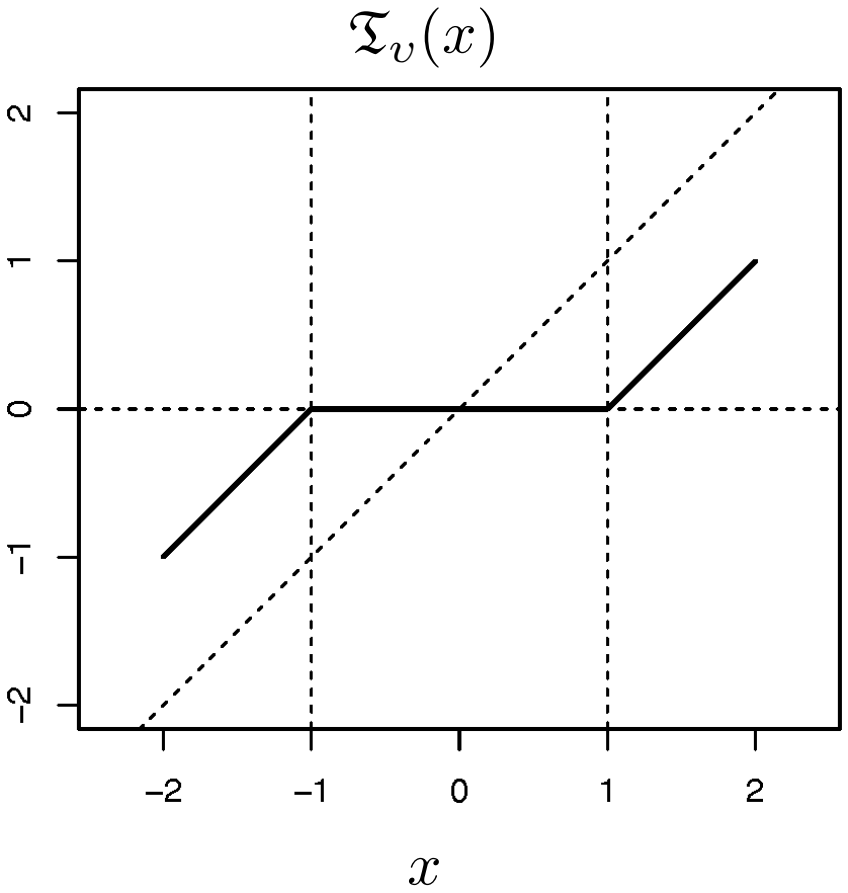
  which is equivalent to $\beta_k(\cdot) = \mathfrak{T}_\upsilon\{\tilde{\beta}_k(\cdot)\}$ with $\tilde{\beta}_k(\cdot) \sim \mathcal{GP}\{0, \kappa(\cdot,\cdot)\}$,
- The soft-thresholding operator:

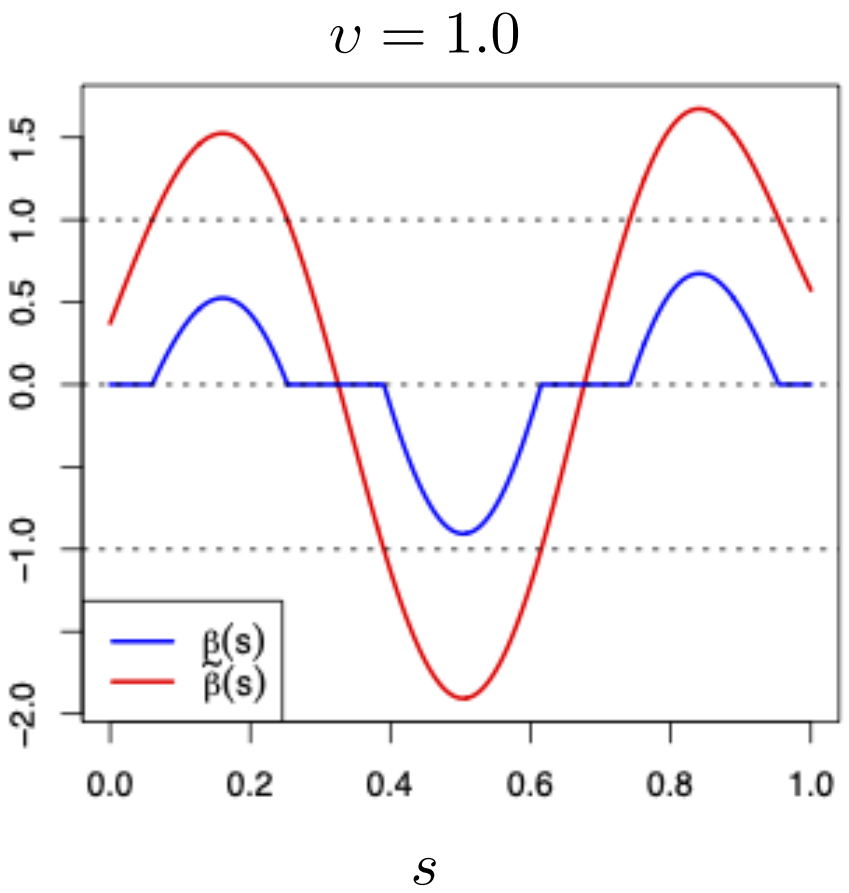
$$\mathfrak{T}_\upsilon(x) = \begin{cases} 0, & |x| \le \upsilon, \\ \mathrm{Sign}(x)(|x| - \upsilon), & |x| > \upsilon, \end{cases}$$

  where $\mathrm{Sign}(x) = 1$ if $x > 0$, $\mathrm{Sign}(x) = -1$ if $x < 0$, $\mathrm{Sign}(0) = 0$.
- The thresholding parameter $\upsilon > 0$ determines the degree of sparsity.

[16] Kang, J. *et al.* Scalar-on-image regression via the soft-thresholded Gaussian process. *Biometrika* **105**, 165–184 (2018).

Illustration: STGP





Theory: Model Identifiability

- Suppose $\boldsymbol{\eta} = \mathbf{0}_K$ and $\boldsymbol{\alpha}$ is known.
- Let $\boldsymbol{\Theta}$ be the parameter space of $\boldsymbol{\zeta}$ and $\boldsymbol{\beta}(\cdot)$.
- Key conditions: activation function $h(\cdot)$ is non-linear and non-quadratic[17].

Proposition 1

Under some regularity conditions, the proposed spatially varying neural network is identifiable up to permutation. Specifically, given a set of locations $\{s_j\}_{j=1}^p \subset \mathcal{B}$, $\{\mathbf{X}_i\}_{i=1}^n \in \mathbb{R}^{n\times p}$, $\{\mathbf{W}_i\}_{i=1}^n \in \mathbb{R}^{n\times q}$ *and* $\boldsymbol{\alpha} \in \mathbb{R}^q$, *if two parameters* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$ *satisfy, for* $i = 1, \ldots, n$, $\pi(Y_i \mid \mathbf{X}_i, \mathbf{W}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}) = \pi\{Y_i \mid \mathbf{X}_i, \mathbf{W}_i, \boldsymbol{\alpha}, \boldsymbol{\theta}'\}$, *for any* $Y_i \in \mathbb{R}$, *then there exists a permutation of* $\{1, \ldots, K\}$, *denoted as* $\{\tau(1), \ldots, \tau(K)\}$ *such that* $\boldsymbol{\theta}' = \{\boldsymbol{\theta}_{\tau(1)}, \ldots, \boldsymbol{\theta}_{\tau(K)}\}$.

[17] Yuan, M. On the identifiability of additive index models. *Statistica Sinica*, 1901–1911 (2011).

Theory: Large Prior Support

Key condition: the true space of the spatially varying weights are adequately smooth

Theorem 1

Under some regularity conditions, for any threshold parameter $\upsilon > 0$ *and any* $\boldsymbol{\theta}^* = \{\boldsymbol{\beta}^*, \boldsymbol{\zeta}^*\} \in \boldsymbol{\Theta}$ *and* $\varepsilon > 0$,

$$Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_\infty < \varepsilon, ||\boldsymbol{\zeta} - \boldsymbol{\zeta}^*||_\infty < \varepsilon) > 0.$$

Remark: the universal approximation probability of neural networks enable the possibility of accurate prediction and feature selection simultaneously.

Theory: Consistency

Theorem 2

Under some conditions, for any $\varepsilon > 0$, *as* $n \to \infty$,

$$Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta} : ||\boldsymbol{\beta} - \boldsymbol{\beta}^*||_1 < \varepsilon, ||\boldsymbol{\zeta} - \boldsymbol{\zeta}^*||_1 < \varepsilon \mid \mathbf{D}_n) \to 1,$$

in $P_{\boldsymbol{\theta}^*}^{(n)}$ *probability.*

Theorem 3

Suppose the model assumptions, prior settings and regularity conditions for Theorem 2 hold, then as $n \to \infty$,

$$Pr(\boldsymbol{\theta} \in \boldsymbol{\Theta} : \text{Sign}\left\{\boldsymbol{\beta}(s)\right\} = \text{Sign}\left\{\boldsymbol{\beta}^*(s)\right\}, s \in \mathcal{B} \mid \mathbf{D}_n) \to 1,$$

in $P_{\boldsymbol{\theta}^*}^{(n)}$ *probability.*

Posterior Computation

- Model representation: using the Karhunen-Loève expansion

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

where $\tilde{\upsilon} = \upsilon / \sigma_b$; eigen functions $\{\varphi_l(s)\}_{l=1}^{\infty}$ and eigen values $\{\lambda_l\}_{l=1}^{\infty}$ are determined by the covariance kernel function $\tilde{\kappa}(s, s')$.

- Modified Squared Exponential (MSE) Kernel:

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

Key appealing property: the eigen decomposition has a closed form, where the eigen functions can be constructed by Hermite polynomials.

Stochastic Gradient Langevin Dynamics

Algorithm 1 Posterior sampling of BNN-STGP

Input: $\mathbf{D}_n = \{(Y_i, \mathbf{X}_i, \mathbf{W}_i)\}_{i=1}^n$, $K, a, b, \tilde{v}, L$,
step size $\{\epsilon_t\}$, minibatch size $|\mathcal{I}|$, number of epochs T,
maximum saves $n_{\max}$, burn-in time t_{burn}.

Output: MCMC samples of parameters in BNN-STGP

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: $\{i\}_{i=1}^n = \bigcup_{j=1}^J \mathcal{I}_j$ with $|\mathcal{I}_j| \leq |\mathcal{I}|$.

    for $j = 1, \cdots, J$ **do**

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

      $\tilde{\boldsymbol{\theta}} \leftarrow \tilde{\boldsymbol{\theta}} + \epsilon_t \nabla \hat{U}_{\mathcal{I}_j}(\tilde{\boldsymbol{\theta}}) + \mathcal{N}(0, 2\epsilon_t)$

    if $t > t_{\text{burn}}$ **then**

      if number of current saves $> n_{\max}$ **then**

        Delete the oldest save of $\tilde{\theta}$.

      Save current value of $\tilde{\boldsymbol{\theta}}$ as one posterior sample.

Bayesian Inference on Region Selection

- MCMC samples: $\{\beta_k^{(t)}(s)\}_{t=1}^T$ for any $s \in \mathcal{B}$.
- Spatially varying posterior inclusion probability of the kth hidden unit:

$$\mathfrak{p}_k(s) = \Pr\{\beta_k(s) \neq 0 \mid \mathbf{D}_n\} \approx \hat{\mathfrak{p}}_k(s) = T^{-1} \sum_{t=1}^{T} I\{\beta_k^{(t)}(s) \neq 0\}$$

- Region selection with Bayesian FDR control[18]

$$\hat{\delta}_k(s) = I\{\hat{\mathfrak{p}}_k(s) \geq \hat{\mathfrak{p}}_{k(j_{\alpha^*})}\}$$

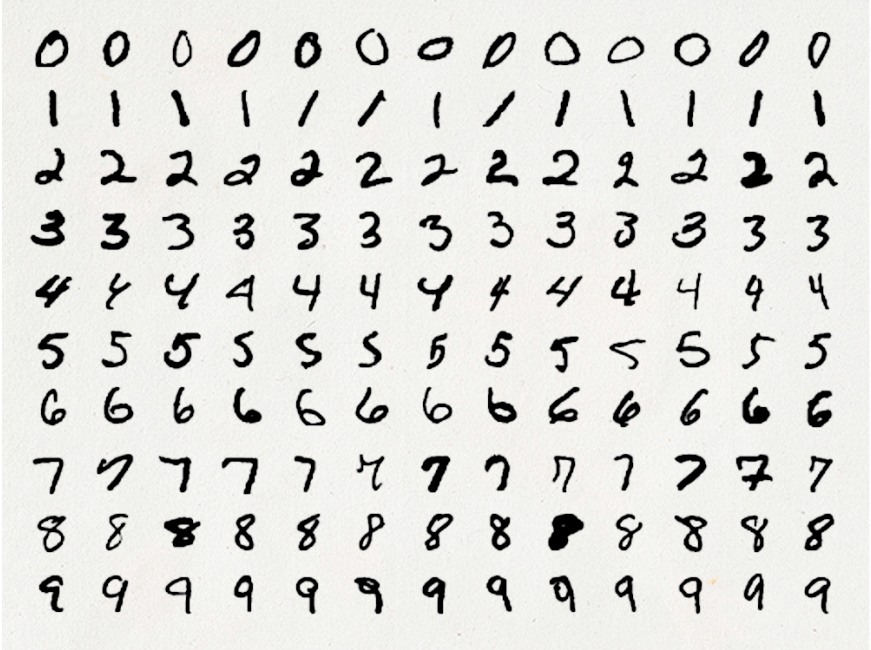
with

$$j_{\alpha^*} = \max\left[l : l^{-1} \sum_{j=1}^{l} \{1 - \hat{\mathfrak{p}}_{k(j)}\} \leq \alpha^*\right].$$

[18] Morris, J. S. *et al.* Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics* **64,** 479–489 (2008).

Application I: MNIST

- The MNIST database of handwritten digits
  - $60,000$ training images
  - $10,000$ test images
- Randomly select $n = 10, 20, 30, 40, 50$ training samples from each digit



- Comparison: BNN-STGP versus BNN[19], CNN[20], DNN
  - BNN-STGP, BNN and DNN, a single hidden layer with 8 hidden nodes.
  - CNN: one convolution layer with 8 filters (kernel $= 3 \times 3$).

[19] Liang, F. *et al.* Bayesian Neural Networks for Selection of Drug Sensitive Genes. *Journal of the American Statistical Association* **0,** 1–18 (2018).

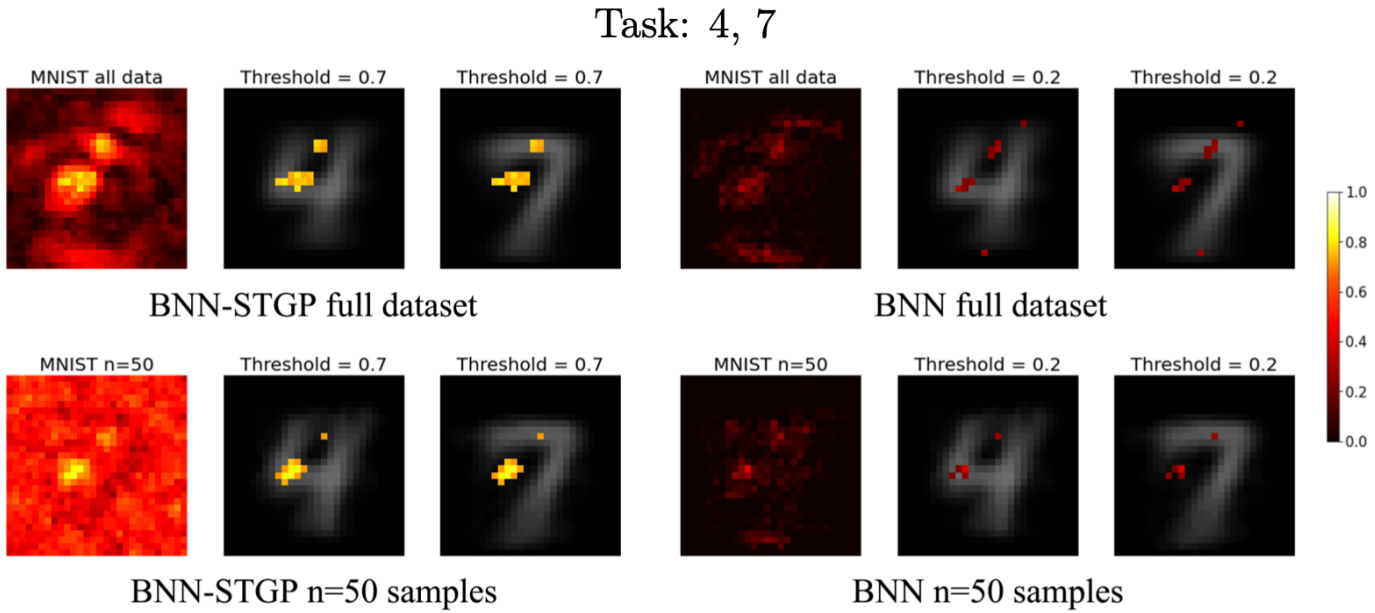
[20] LeCun, Y. *et al.* Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems* **2** (1989).

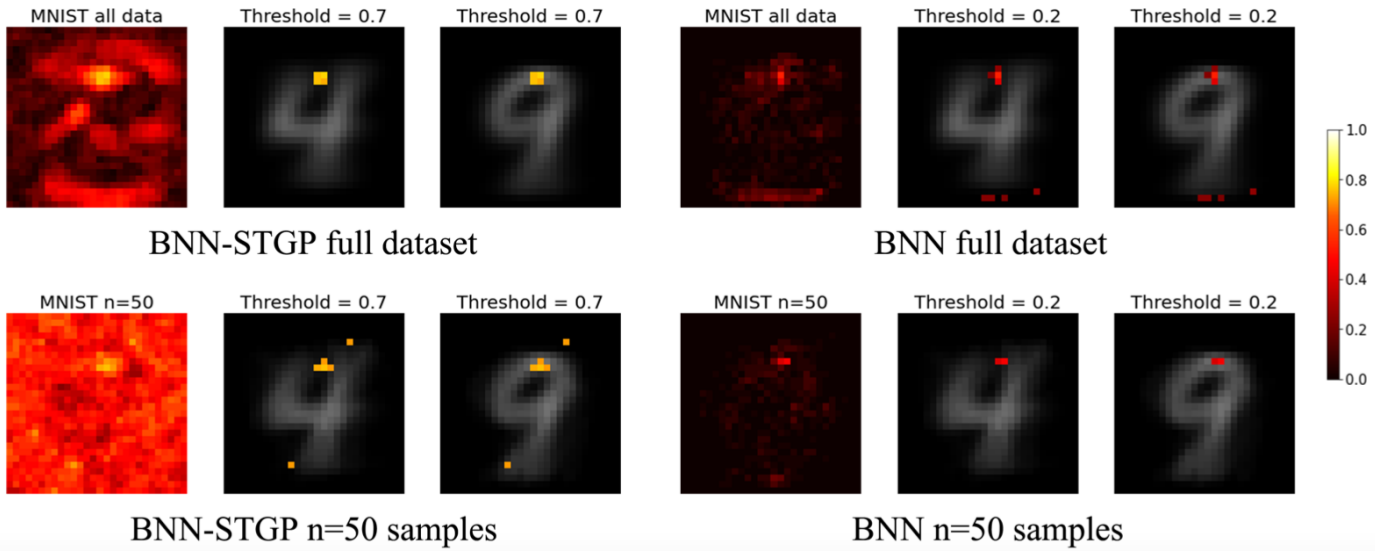
Classification with small training sample size

Table 1: Binary Classification Error Rate(%) on MNIST Data

Method	Task	n=10	n=20	n=30	n=40	n=50	All
BNN-STGP	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

Image feature selection stability





Application II: Analysis of ABCD task fMRI data

- The Adolescent Brain Cognitive Development (ABCD) Study[21]
  - The largest long-term study of child brain development in US
  - Collect rich data sets from 21 research sites across the country (11,878 children ages 9-10 to join the study).
- Questions:
  - Can we predict general intelligence from the task fMRI?
  - Which brain regions are important for prediction?
- Outcome variable: general cognitive ability, i.e., g-factor
- Imaging predictors: contrast maps of task fMRI (n-back), about 2,000 subjects, 180,000 voxels from 90 AAL[22] regions
  - 2-back versus 0-back; 2-back versus baseline; 0-back versus baseline
- Non-image predictors: p-factor, Age, gender, Race, Household Marital Status, Household Income,

[21] Casey, B. *et al.* The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites. *Developmental cognitive neuroscience* **32,** 43–54 (2018).

[22] Tzourio-Mazoyer, N. *et al.* Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15,** 273–289 (2002).

Model settings

- Baseline model: only use non-image predictor
- BNN-STGP: single layer neural network with $K = 20$ hidden units
- Consider Two CNN models with a similar network architecture

  Input - Conv3d - MaxPool3d - Conv3d -MaxPool3d - fully-connected - Output.

- CNN1: 8 filters (kernel $= 3 \times 3$, stride $= 2$) and 16 filters (kernel $= 4 \times 4$, stride $= 2$) with directly connecting the output.
- CNN2: 16 filters and 32 filters in the convolutional layers followed by a fully connected layer with 20 hidden units before the output.

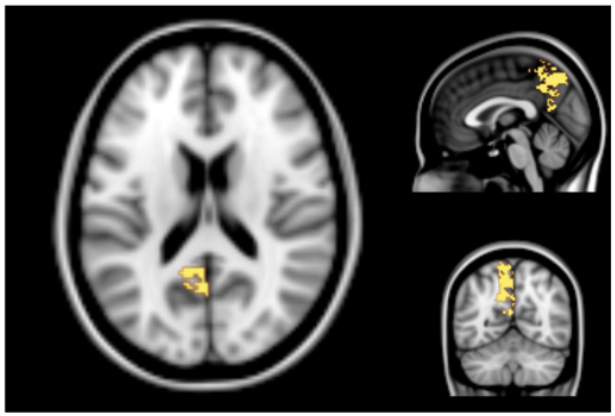
Predictive R-squared via Cross Validations

- Random-split: training 80% and test 20% for 100 times.
- Single-site: training one site (over 100 children), and test 20 sites.

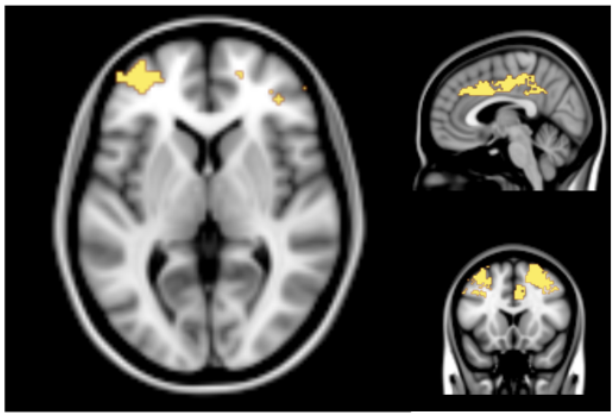
Modality	Method	Random split	Single site
No Image	Linear Regression	0.243(0.032)	0.150(0.055)
2 back-0 back	BNN-STGP	**0.320(0.014)**	**0.195(0.041)**
	CNN1(8-16)	0.307(0.028)	0.186(0.053)
	CNN2(16-32-20)	0.311(0.025)	0.182(0.047)
2 back-baseline	BNN-STGP	**0.301(0.020)**	**0.184(0.046)**
	CNN1(8-16)	0.289(0.034)	0.177(0.056)
	CNN2(16-32-20)	0.288(0.036)	0.179(0.045)
0 back-baseline	BNN-STGP	0.284(0.012)	**0.166(0.057)**
	CNN1(8-16)	0.272(0.029)	0.159(0.068)
	CNN2(16-32-20)	**0.286(0.027)**	0.164(0.048)

Region Selection: 2 back - 0 back Contrast Maps

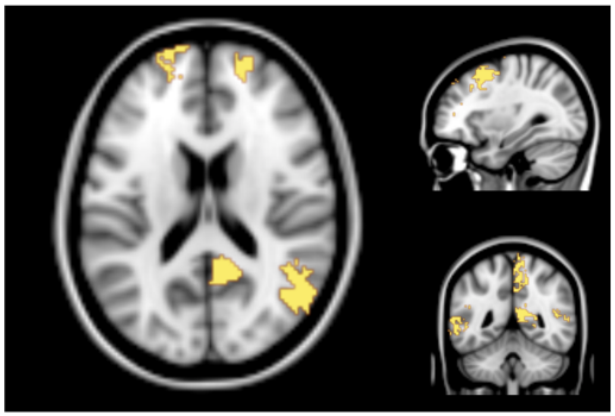
Region Code	Region Name	# Selected Voxels
6301	Precuneus_L	1471
6302	Precuneus_R	1373
2201	Frontal_Mid_L	1317
4012	Cingulum_Mid_R	1170
2101	Frontal_Sup_L	1062
4011	Cingulum_Mid_L	1056
2102	Frontal_Sup_R	981
2202	Frontal_Mid_R	897
8202	Temporal_Mid_L	851
5022	Lingual_R	828



(a) Memory Retrieval



(b) Task Control



(c) Default Mode

Summary

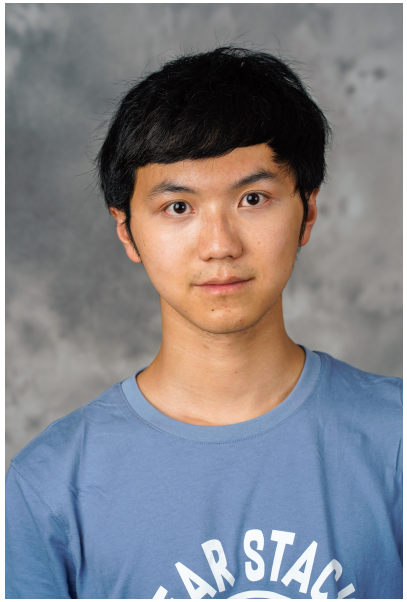
BNN-STGP

- A novel Bayesian neural network for scalar-on-image regression
- STGP prior: sparse and spatially varying weight parameters
- Theoretical properties: model identifiability, posterior consistency
- Posterior computation: scalable and efficient SGLD algorithm with the MSE kernel expansion.
- Performance:
  - High prediction accuracy for small training sample
  - Stable feature/region selection
  - Interpretable results in computer vision and neuroimaging applications.
- Key references
  - Kang J, Reich BJ, Staicu AM (2018) Scalar-on-image regression via the soft thresholded Gaussian process. *Biometrika*: 105(1) 165–184.
  - Wu, Wu and Kang (2021) Bayesian spatially-varying weight neural networks via the soft thresholded Gaussian process: Under Review.

Acknowledgement



Ben Wu
Renmin University of China



Keru Wu
Duke University

- Grant Support: NIH R01DA048993, R01MH105561, R01GM124061 and NSF IIS2123777

Thank you and Questions?