

# Supplement for “Subject-Specific Scalar-on-Image Regression”

LEO YU-FENG LIU<sup>1,†</sup>, HAIXU MA<sup>1,†</sup>, YUFENG LIU<sup>3,\*</sup>, AND HONGTU ZHU<sup>1,2</sup>

<sup>1</sup>*Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, USA*

<sup>2</sup>*Department of Biostatistics, University of North Carolina at Chapel Hill, USA*

<sup>3</sup>*Department of Statistics, University of Michigan, USA*

We discuss the detailed algorithms for the estimation procedure, model selection, and the prediction. Note that in the supplementary materials, we use the prefix “S” for equation numbers (e.g., equation (S.1)) and figure labels (e.g., Figure S1) to distinguish them from those in the main paper. The code implementation is available in the supplementary materials.

## S1 Estimation and prediction

We discuss how to estimate the unknown coefficients in the model and establish the prediction rules. In particular, the population disease map  $\beta$  is estimated in the functional regression model. The distribution of the homogeneous masking image  $\mathbf{B}$  is obtained by using the maximum of the posterior probability. The individual-level parameter  $\mathbf{B}_i = \sum_{k=1}^K I_{ik} \mathbf{R}_k$  is determined by the estimated individual assignment  $I_{ik}$ . The estimation of homogeneous and heterogeneous coefficients can be obtained by an EM-type of algorithm. In addition, the prediction rule is constructed via a pattern matching process.

### S1.1 Homogeneous disease map $\beta$

In our S3IR model, the heterogeneity is characterized by the subject-specific masking image  $\mathbf{B}_i$ ’s as stated in Equation (4). Given the binary image  $\mathbf{B}_i$ ’s, we can proceed the feature screening procedure and extract the informative regions in each covariate image described by  $\mathbf{B}_i$ ’s. The masked covariate images are defined as  $\tilde{\mathbf{X}}_i = \mathbf{B}_i \circ \mathbf{X}_i$ . Then the regression model (4) can be rewritten with the following form,

$$y_i = \beta_0 + \int_{t \in \mathcal{D}} \tilde{\mathbf{X}}_i(t) \beta(t) dt + \epsilon_i. \quad (\text{S.1})$$

Note that model (S.1) is a homogeneous functional regression defined in Equation (1), with covariates  $\tilde{\mathbf{X}}_i$ ’s and response  $y_i$ ’s. To simplify the notation, we further center all the covariates and responses so that the intercept term  $\beta_0$  can be dropped in the derivation. Since the coefficient image  $\beta$  lies in a RKHS induced by the radial based kernel  $\mathcal{K}$  given by Equation (6), we can estimate the coefficient image by using the kernel ridge regression method in Yuan and Cai (2010). In particular, the parameters can be estimated by solving the following empirical risk minimization,

$$\hat{\beta} = \underset{\beta \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left( y_i - \int_{t \in \mathcal{D}} \tilde{\mathbf{X}}_i(t) \beta(t) dt \right)^2 + \lambda_n J(\beta), \quad (\text{S.2})$$

<sup>†</sup>Contributed equally.

\*Corresponding author. Email: [yufliu@umich.edu](mailto:yufliu@umich.edu).

where the first term is the squared loss which ensures the goodness of fit on the training data, and the second term  $J(\beta)$  represents the smoothness penalty. Here,  $\lambda_n$  is the tuning parameter that controls the level of the smoothness and the complexity of the estimated coefficients.

According to the representer theorem (Wahba, 1990), there exists a parameter  $c \in \mathbb{R}^n$  such that the solution to (S.2) can be expressed as follows,

$$\hat{\beta}(t) = \sum_{i=1}^n c_i \int_{s \in \mathcal{D}} \mathcal{K}(t, s) \tilde{\mathbf{X}}_i(s) ds, \quad (\text{S.3})$$

where  $c_i$  represents the  $i$ -th component of the parameter vector  $c$ .

The penalty can be written as  $J(\beta) = c' \Sigma c$ , where  $\Sigma$  is a  $n$  by  $n$  Gram matrix with  $\Sigma_{ij} = \iint \tilde{\mathbf{X}}_i(s) \mathcal{K}(s, t) \tilde{\mathbf{X}}_j(t) ds dt$ , for  $i, j \in \{1, \dots, n\}$ . Therefore, the empirical risk minimization in (S.2) is equivalent to the following problem:

$$\begin{aligned} \hat{c} &= \underset{c \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left( y_i - \int_{t \in \mathcal{D}} \tilde{\mathbf{X}}_i(t) \sum_{j=1}^n c_j \int_{s \in \mathcal{D}} \mathcal{K}(t, s) \tilde{\mathbf{X}}_j(s) ds dt \right)^2 + \lambda_n c' \Sigma c \\ &= \underset{c \in \mathbb{R}^n}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n \left( y_i - \sum_{j=1}^n c_j \iint \tilde{\mathbf{X}}_i(t) \mathcal{K}(t, s) \tilde{\mathbf{X}}_j(s) ds dt \right)^2 + \lambda_n c' \Sigma c \\ &= \underset{c \in \mathbb{R}^n}{\operatorname{argmin}} \|Y - \Sigma c\|^2 + n \lambda_n c' \Sigma c, \end{aligned}$$

where  $Y = (y_1, \dots, y_n)'$  denotes the response vector. The above problem is referred as the Tikhonov regularization problem (Tikhonov et al., 1943), which has a close form solution for  $\hat{c}$ , i.e.,  $\hat{c} = (\Sigma^2 + n \lambda \Sigma)^{-1} \Sigma Y$ . Finally, the coefficient image  $\beta$  is recovered by plugging  $\hat{c}$  into (S.3):

$$\hat{\beta}(t) = \sum_{i=1}^n \hat{c}_i \int_{s \in \mathcal{D}} \mathcal{K}(t, s) \tilde{\mathbf{X}}_i(s) ds. \quad (\text{S.4})$$

## S1.2 Homogeneous region detection

The population-level masking image  $\mathbf{B}$  follows the Potts model with the probability mass function specified in Equation (7). In order to estimate  $\mathbf{B}$ , we implement the method of maximizing the posterior probability (Bassett and Deride, 2019). The estimator  $\hat{\mathbf{B}}$  is obtained by solving the following optimization function of the joint likelihood of  $Y$  and  $\mathbf{B}$ :

$$\begin{aligned} \hat{\mathbf{B}} &= \underset{\mathbf{B} \in \mathcal{D}_{\mathbf{B}}}{\operatorname{argmax}} L(Y, \mathbf{B}; \hat{\beta}, \tilde{\mathbf{X}}_i \text{'s}, \tau, \sigma^2) \\ &= \underset{\mathbf{B} \in \mathcal{D}_{\mathbf{B}}}{\operatorname{argmax}} \prod_{i=1}^n \left( \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{r_i^2}{2\sigma^2} \right\} \right) \cdot \exp \left\{ \tau \sum_{t \in \mathcal{D}} s_{\mathbf{B}}(t) \right\} C(\tau) \\ &= \underset{\mathbf{B} \in \mathcal{D}_{\mathbf{B}}}{\operatorname{argmin}} \sum_{i=1}^n r_i^2 - \tau \sum_{t \in \mathcal{D}} s_{\mathbf{B}}(t), \end{aligned} \quad (\text{S.5})$$

where  $\mathcal{D}_{\mathbf{B}} = \{0, 1\}^{\otimes |\mathcal{D}|}$  represents the set of all possible binary images on  $\mathcal{D}$ ,  $r_i = y_i - \int_{t \in \mathcal{D}} \tilde{\mathbf{X}}_i(t) \mathbf{B}(t) \hat{\beta}(t) dt$  corresponds to the residual term and  $\sigma^2$  represent the variance of the prediction.

Note that the above optimization problem is a non-convex integer programming problem and the feasible set  $\mathcal{D}_{\mathbf{B}} = \{0,1\}^{\otimes |\mathcal{D}|}$  is finite. One approach to solve this is to enumerate all possible elements in  $\mathcal{D}_{\mathbf{B}}$  at the cost of  $\mathcal{O}(2^{|\mathcal{D}|})$  operations. This is in general unachievable when the size of the image is large. In order to solve this computational problem, we adapt the Iterative Conditional Modes algorithm proposed by Besag (1986), which uses a greedy iterative strategy to search for a local minimum. By utilizing this algorithm, the convergence of the algorithm is usually achieved after a few iterations with a complexity of  $\mathcal{O}(|\mathcal{D}|)$  operations.

By considering the Potts prior for the distribution of  $\mathbf{B}$ ,  $\hat{\mathbf{B}}$  usually yields a pattern with disjoint regions. We label those regions as  $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_K$ , and decompose  $\hat{\mathbf{B}}$  into  $K$  binary images  $\hat{\mathbf{R}}_k$ 's as defined in Equation (11). In addition, we introduce a parameter  $\tau$  that controls the level of spatial smoothness and local similarity. The maximum likelihood estimation is generally difficult to compute due to the intractable normalization factor  $C(\tau)$ , which requires summing over all  $2^{|\mathcal{D}|}$  possible binary configurations. To circumvent this computational challenge, we use the Pseudo-Likelihood (PL) approach, which approximates the full likelihood by treating each pixel's value as conditionally independent given its neighbors, thereby avoiding the need to compute the intractable normalization constant:

$$\begin{aligned} L(\hat{\mathbf{B}}; \tau) &= \prod_{t \in \mathcal{D}} \text{PL}(\hat{\mathbf{B}}(t) | \hat{\mathbf{B}}) \\ &= \prod_{t \in \mathcal{D}} \frac{\exp\left(\tau \sum_{s \in \mathcal{N}(t)} \delta(\hat{\mathbf{B}}(s), \hat{\mathbf{B}}(t))\right)}{\exp\left(\tau \sum_{s \in \mathcal{N}(t)} \delta(\hat{\mathbf{B}}(s), 1)\right) + \exp\left(\tau \sum_{s \in \mathcal{N}(t)} \delta(\hat{\mathbf{B}}(s), 0)\right)}. \end{aligned}$$

For any given  $\hat{\mathbf{B}}$ , the parameter  $\hat{\tau}$  is estimated by setting the derivative of PL to zero, i.e.,  $\partial \ln(\text{PL}(\hat{\mathbf{B}}; \tau)) / \partial \tau = 0$ . These two steps would alternate with each other until the final convergence.

### S1.3 Heterogeneous regions assignments

We introduce the estimation of the heterogeneous coefficients. As defined in Equation (4), our model characterizes the heterogeneity through the individual-level binary masking image  $\mathbf{B}_i = \sum_{k=1}^K I_{ik} \mathbf{R}_k$ . Given the estimation of the homogeneous masking image  $\hat{\mathbf{B}}$  described in Section S1.2, it suffices to estimate the indicators  $I_{ik}$ 's. We assume that the region assignment for each subject are independent, and thus they can be estimated component-wisely by minimizing the prediction squared error. Given the estimated  $\hat{\boldsymbol{\beta}}$  and  $\hat{\mathbf{B}}$ ,  $I_{ik}$ 's are estimated as follows:

$$\begin{aligned} [\hat{I}_{i1}, \dots, \hat{I}_{iK}] &= \underset{\{0,1\}^{\otimes K}}{\text{argmin}} \left( y_i - \sum_{k=1}^K I_{ik} \int \mathbf{X}_i(t) \hat{\mathbf{R}}_k(t) \hat{\boldsymbol{\beta}}(t) dt \right)^2 \\ &= \underset{\{0,1\}^{\otimes K}}{\text{argmin}} \left| y_i - \sum_{k=1}^K I_{ik} \hat{\mu}_{ik} \right|, \end{aligned} \tag{S.6}$$

where  $\hat{\mu}_{ik} = \int \mathbf{X}_i(t) \hat{\mathbf{R}}_k(t) \hat{\boldsymbol{\beta}}(t) dt$ . Problem (S.6) is a binary integer programming problem. Due to the Potts prior, the total number of regions  $K$  detected by  $\hat{\mathbf{B}}$  cannot be too large. Thus, we can consider the enumerative search on the feasible set of  $\{0,1\}^{\otimes K}$  to find the best solution. The optimization for each subject can be efficiently conducted using parallel computing. In practice, the algorithm may overfit by assigning false positive regions with weak signal, i.e.,  $|\hat{\mu}_{ik}|$  is small.

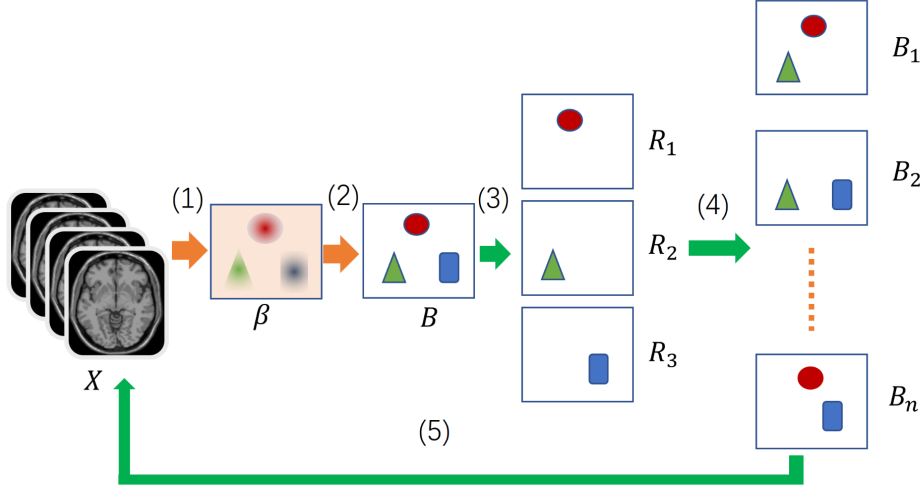


Figure S1: The flow chart of the estimation procedure.

In order to avoid this issue, we impose a penalty on  $I_{ik}$ 's, which leads to the following optimization problem,

$$[\hat{I}_{i1}, \dots, \hat{I}_{iK}] = \underset{\{0,1\}^{\otimes K}}{\operatorname{argmin}} \left| y_i - \sum_{k=1}^K I_{ik} \hat{\mu}_{ik} \right| + \lambda_s \sum_{k=1}^K |I_{ik}|, \quad (\text{S.7})$$

where  $\lambda_s$  is a parameter controls the sparsity level of the region assignment. In particular, by imposing such a penalty, the value of the optimization function would decrease  $\lambda_s$  when assigning an associated region as active. In our implementation,  $\lambda_s$  is determined according to the variation of training samples, i.e.,  $\lambda_s = 0.05\text{Var}(Y)$ .

In summary, the whole estimation procedure can be illustrated by the flowchart in Figure S1. The orange arrow characterizes the estimation procedure for the population-level coefficient  $\beta$  and  $B$ ; the green arrow refers to the individual-level estimation. In particular, based on the masked covariate image  $\tilde{X}_i$ 's, we estimate  $\beta$  in Step (1) according to Equation (S.4), and then update  $B$  in Step (2) by maximizing the joint likelihood defined in Equation (S.5). In Step (3),  $R_k$  is extracted based on the estimation of  $B$ . These steps only estimate the population-level coefficients and are not related to the individual-level region assignments. In Step (4), the region assignments  $I_{ik}$ 's are obtained by minimizing the penalized training loss according to Equation (S.6). Based on the region assignments, we further update the heterogeneous hidden layer  $B_i$ 's, and reconstruct the covariates  $\tilde{X}_i$ 's in Step (5). The whole estimation algorithm is conducted iteratively with an EM-type estimation procedure until the convergence criteria is met. In this paper, we terminate the algorithm when the difference of individual-level coefficient  $\beta_i$ 's between two adjacent iterations is less than a prespecified threshold.

#### S1.4 Tuning parameter selection

The population-level of our model includes kernel ridge regression estimation on the RKHS. There are two tuning parameters involved: the bandwidth  $\sigma$  for the radial based kernel in Equation (6), and the ridge penalty level  $\lambda_n$  in Equation (S.2). In this paper, we use the Bayesian Information

Criteria (BIC) as the guidance to select the parameters. In particular, according to the Stein's Unbiased Risk Estimation (SURE) theory (Stein, 1981), the degree of freedom is defined as  $df = \sum_{i=1}^n \text{cov}(\hat{\mu}_i, y_i) = \sum_{i=1}^n \partial \hat{\mu}_i / \partial y_i$ . In our model, the estimation step for  $\beta$  can be expressed as:  $\hat{Y} = \Sigma(\Sigma^2 + n\lambda\Sigma)^{-1}\Sigma Y := HY$ , where  $\Sigma_{i,j} = \iint \tilde{\mathbf{X}}_i(s)\mathcal{K}(s,t)\tilde{\mathbf{X}}_j(t)dsdt$ ,  $\tilde{\mathbf{X}}_i = \mathbf{X}_i \circ \mathbf{B}_i$  and  $H = \Sigma(\Sigma^2 + n\lambda\Sigma)^{-1}\Sigma$ . Thus, the degree of freedom can be approximated by  $df = \text{trace}(H)$ , and the BIC is therefore given by  $\text{BIC} = \log(n) \times df + n \log(\text{RSS}/n)$ , where  $\text{RSS} = \sum_{i=1}^n (\hat{y}_i - y)^2$  represents the residual sums of square. The parameters  $\sigma$  and  $\lambda$  are chosen to minimize the BIC.

Since the estimation of  $\hat{\mathbf{B}}$  and  $\hat{\mathbf{B}}_i$ 's do not involve tuning, we would not select the parameters after the entire estimation process. Hence, we only embed the tuning procedure in the kernel ridge regression, i.e., Step (2) in Figure S1. Therefore, the entire estimation procedure can be viewed as a tuning free algorithm.

## S1.5 Prediction

The prediction for the models involving heterogeneous structure is a challenging problem. In order to predict the response  $y^*$  for a new subject with covariate  $\mathbf{X}^*$ , the mixture regression models often consider the posterior probability  $\hat{\pi}_k$  and utilize the weighted pooling, i.e.,  $\hat{y}^* = \sum_{k=1}^K \hat{\pi}_k \int_{t \in \mathcal{D}} \mathbf{X}^*(t) \hat{\beta}_k(t) dt$ . However, this estimation approach often leads to suboptimal prediction accuracy at the individual level, as it overlooks the heterogeneous information regarding the relationship between  $\mathbf{X}$  and  $y$  by assuming the identical distribution for all subjects. To address this issue, a joint mixture regression model was proposed by Hoshikawa (2013), which incorporates the joint distribution between the covariates  $\mathbf{X}$  and the response  $y$  for prediction. Specifically, a strategy known as covariate-specific pooling is employed, i.e.,  $\hat{y}^* = \sum_{k=1}^K \hat{\pi}_k(\mathbf{X}^*) \int_{t \in \mathcal{D}} \mathbf{X}^*(t) \hat{\beta}_k(t) dt$ , where  $\hat{\pi}_k(\mathbf{X}^*)$  denotes the estimated probability of  $\mathbf{X}^*$  in group  $k$ . The enhancement in prediction accuracy relies on the Kullback-Leibler divergence of the component-wise distribution.

In contrast to the utilization of probabilistic distributions in mixture regression models, the heterogeneity in our proposed S3IR model is characterized by the subject-specific masking image. For a new subject with covariate  $\mathbf{X}^*$ , we use  $\mathbf{B}_{\mathbf{X}^*}$  to denote the subject-specific masking image. To predict the response for a new subject, we first determine its region assignment, and then construct the corresponding individual-level coefficient  $\hat{\mathbf{B}}_{\mathbf{X}^*}$ . Finally, the predicted response is given by  $\hat{y}^* = \int_{t \in \mathcal{D}} \mathbf{X}^*(t) \hat{\mathbf{B}}_{\mathbf{X}^*}(t) \hat{\beta}(t) dt$ .

Note that in the proposed S3IR model, the homogeneous disease map  $\beta$  serves to screen out the irrelevant regions at the population level. Additionally, the hidden binary image  $\mathbf{B}_{\mathbf{X}^*}$  helps to further select individual-level active regions for each subject with future covariate  $\mathbf{X}^*$ . Hence, the regression model is able to enhance the prediction performance through individualized signal detection. The estimation of  $\mathbf{B}_{\mathbf{X}^*}$  proceeds as follows. Based on the estimated individual region assignments  $I_{ik}$  for training data discussed in Section S1.3, for each detected region  $\mathcal{R}_k$ , we compute the component-wise average of  $\mathbf{X}_i$ 's in  $\mathcal{R}_k$  with active signals (lesions tissues) when  $I_{ik} = 1$ , and inactive signals (normal tissues) when  $I_{ik} = 0$ . In particular, for each  $t \in \mathcal{D}$ , the component-wise average of  $\mathbf{X}_i$ 's in  $\mathcal{R}_k$  is computed as follows,

$$\begin{aligned} \text{Average for Active Regions: } \mathbf{P}_{k1}(t) &= \frac{\sum_{i=1}^n I_{ik} \mathbf{X}_i(t) \circ \mathbf{R}_k}{\sum_{i=1}^n I_{ik}}, \\ \text{Average for Inactive Regions: } \mathbf{P}_{k0}(t) &= \frac{\sum_{i=1}^n (1 - I_{ik}) \mathbf{X}_i(t) \circ \mathbf{R}_k}{\sum_{i=1}^n (1 - I_{ik})}. \end{aligned}$$

Based on above averages computed from the training data, we construct a classifier for each  $\mathcal{R}_k$ , which is used to determine whether the region  $\mathcal{R}_k$  is active or not for a new subject. The classification rule is determined by measuring the distance between  $\mathbf{X}^*$  and the average of  $\mathbf{X}_i$ 's such as  $\mathbf{P}_{k0}$  and  $\mathbf{P}_{k1}$  in active and inactive regions:

$$\hat{I}_k^* = \mathbf{1} \{d(\mathbf{X}^* \circ \mathbf{R}_k, \mathbf{P}_{k1}) < d(\mathbf{X}^* \circ \mathbf{R}_k, \mathbf{P}_{k0})\}, \quad (\text{S.8})$$

where  $d(\cdot, \cdot)$  measures the weighted distance between  $\mathbf{X}^*$  and the average. Here, a shape-based weighted mean of the squared pixel/voxel-wise difference is employed to compute the distance:

$$d(\mathbf{X}^* \circ \mathbf{R}_k, \mathbf{P}_{kl}) = \sqrt{\int_{t \in \mathcal{R}_k} (\mathbf{X}^*(t) - \mathbf{P}_{kl}(t))^2 W(t) dt}, \quad \text{for } l = 1, 0,$$

where  $W(t)$  is the weight assigned to each pixel  $t$ , with larger weights assigned to the pixels near the center of the region  $\mathcal{R}_k$ , and smaller weights assigned to the pixels near the boundary. We introduce the weight to consider different magnitudes for various pixels with the center of  $\mathcal{R}_k$  given the largest magnitude. This weighting strategy accounts for the fact that the estimated region  $\mathcal{R}_k$  from Section S1.2 may have a higher probability of covering the center of the true region, but may not be able to detect the boundary precisely. In our empirical studies, this weighting strategy can effectively reduce the classification error. The shape-based weights  $W(t)$  are illustrated in Figure S2. For a given region  $\mathcal{R}_k$  in the left panel, the center of  $\mathcal{R}_k$  (the red pixel shown in the right panel) is assigned with the largest weight  $W_3$  while the pixels on the boundary (the yellow pixels shown in the right panel) are given smaller weight  $W_1$ .

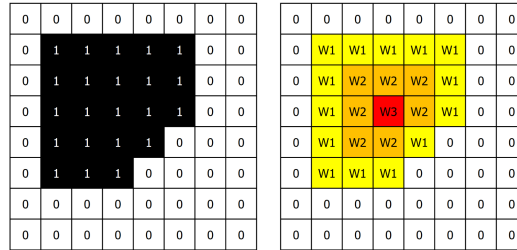


Figure S2: An example of the shape-based weights. The left panel is a binary image, and the right panel is the associated weight matrix, with  $W_1 < W_2 < W_3$ .

Once the classification rule is constructed, the response can be predicted as follows:

$$\hat{y}^* = \sum_{k=1}^K \hat{I}_k^* \int_{t \in \mathcal{R}_k} \mathbf{X}^*(t) \hat{\beta}(t) dt.$$

Note that while our framework focuses on point prediction, uncertainty quantification for  $\hat{y}^*$  can be incorporated through resampling methods such as the bootstrap, or by deriving asymptotic variance estimates under regularity conditions. Specifically, the variability in  $\hat{y}^*$  stems from both the classification  $\hat{I}_k^*$  and the estimated coefficients  $\hat{\beta}(t)$ . Approximating the distribution of  $\hat{y}^*$  enables the construction of confidence intervals, which helps to provide more informative predictions and aid downstream inference.

## References

- Bassett R, Deride J (2019). Maximum a posteriori estimators as a limit of bayes estimators. *Mathematical Programming*, 174(1): 129–144.
- Besag J (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 259–302.
- Hoshikawa T (2013). Mixture regression for observational data, with application to functional regression models. *arXiv preprint arXiv:1307.0170*.
- Stein CM (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 1135–1151.
- Tikhonov AN, et al. (1943). On the stability of inverse problems. In: *Dokl. akad. nauk sssr*, volume 39, 195–198.
- Wahba G (1990). *Spline models for observational data*. Society for Industrial and Applied Mathematics.
- Yuan M, Cai TT (2010). A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6): 3412–3444.