Group Additive Structure Identification for Kernel Nonparametric Regression

Pan Chao
Department of Statistics
Purdue University
West Lafayette, IN 47906
panchao25@gmail.com

Michael Zhu
Department of Statistics, Purdue University
West Lafayette, IN 47906
Center for Statistical Science
Department of Industrial Engineering
Tsinghua University, Beijing, China
yuzhu@purdue.edu

Abstract

The additive model is one of the most popularly used models for high dimensional nonparametric regression analysis. However, its main drawback is that it neglects possible interactions between predictor variables. In this paper, we reexamine the group additive model proposed in the literature, and rigorously define the intrinsic group additive structure for the relationship between the response variable Y and the predictor vector $\boldsymbol{X}$, and further develop an effective structure-penalized kernel method for simultaneous identification of the intrinsic group additive structure and nonparametric function estimation. The method utilizes a novel complexity measure we derive for group additive structures. We show that the proposed method is consistent in identifying the intrinsic group additive structure. Simulation study and real data applications demonstrate the effectiveness of the proposed method as a general tool for high dimensional nonparametric regression.

1 Introduction

Regression analysis is popularly used to study the relationship between a response variable Y and a vector of predictor variables $\boldsymbol{X}$. Linear and logistic regression analysis are arguably two most popularly used regression tools in practice, and both postulate explicit parametric models on $f(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$ as a function of $\boldsymbol{X}$. When no parametric models can be imposed, nonparametric regression analysis can instead be performed. On one hand, nonparametric regression analysis is flexible and not susceptible to model mis-specification, whereas on the other hand, it suffers from a number of well-known drawbacks especially in high dimensional settings. Firstly, the asymptotic error rate of nonparametric regression deteriorates quickly as the dimension of $\boldsymbol{X}$ increases. [23] shows that with some regularity conditions, the optimal asymptotic error rate for estimating a d-times differentiable function is $\mathcal{O}\left(n^{-d/(2d+p)}\right)$, where p is the dimensionality of $\boldsymbol{X}$. Secondly, the resulting fitted nonparametric function is often complicated and difficult to interpret.

To overcome the drawbacks of high dimensional nonparametric regression, one popularly used approach is to impose the additive structure [8] on $f(\boldsymbol{X})$, that is to assume that $f(\boldsymbol{X}) = f_1(X_1) + \cdots + f_p(X_p)$ where $f_1, \ldots, f_p$ are p unspecified univariate functions. Thanks to the additive structure, the nonparametric estimation of f or equivalently the individual f_i's for $1 \leq i \leq p$ becomes efficient and does not suffer from the curse of dimensionality. Furthermore, the interpretability of the resulting model has also been much improved.

The key drawback of the additive model is that it does not assume interactions between the predictor variables. To address this limitation, functional ANOVA models were proposed to accommodate higher order interactions, see [7] and [19]. For example, by neglecting interactions of

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

order higher than 2, the functional ANOVA model can be written as $f(\boldsymbol{X}) = \sum_{i=1}^{p} f_i(X_i) + \sum_{1\le i,j\le p} f_{ij}(X_i, X_j)$, with some marginal constraints. Another higher order interaction model, $f(\boldsymbol{X}) = \sum_{d=1}^{D} \sum_{1\le i_1,\ldots,i_d\le p} f_j(X_{i_1}, \ldots, X_{i_d})$, is proposed by [9]. This model considers all interactions of order up to D, which is estimated by Kernel Ridge Regression (KRR) [16] with the elementary symmetric polynomial (ESP) kernel.

Both of the two models discussed above assume the existence of possible interactions between any two or more predictor variables. This can lead to a serious problem, that is, the number of nonparametric functions that need to be estimated quickly increases as the number of predictor variables increases. To control the explosion of interaction terms, one approach is to impose the sparsity assumption and then use variable selection methods such as the lasso to select only the important interactions. For the functional ANOVA model, the COSSO method developed by [14] followed this approach. [2] proposes hierarchical kernel learning which assumes that the kernel of inputs is decomposable sum of many basis kernels. Then kernel selection is performed to only select important interactions by imposing group lasso type penalty.

There exists another direction to generalize the additive model. When proposing the Optimal Kernel Group Transformation (OKGT) method for nonparametric regression, [17] considers the additive structure of predictor variables in groups instead of individual predictor variables. Let $G := \{\boldsymbol{u}_j\}_{j=1}^{d}$ be a index partition of the predictor variables, that is, $\boldsymbol{u}_j \cap \boldsymbol{u}_k = \emptyset$ if $j \neq k$ and $\cup_{j=1}^{d} \boldsymbol{u}_j = \{1, \ldots, p\}$. Let $\boldsymbol{X}_{\boldsymbol{u}_j} = \{X_k; k \in \boldsymbol{u}_j\}$ for $j = 1, \ldots, d$. Then $\{X_1, \ldots, X_d\} = \boldsymbol{X}_{\boldsymbol{u}_1} \cup \cdots \cup \boldsymbol{X}_{\boldsymbol{u}_d}$. For any function $f(\boldsymbol{X})$, if there exists an index partition $G = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$ such that

$$f(\boldsymbol{X}) = f_{\boldsymbol{u}_1}(\boldsymbol{X}_{\boldsymbol{u}_1}) + \ldots + f_{\boldsymbol{u}_d}(\boldsymbol{X}_{\boldsymbol{u}_d}), \tag{1}$$

where $f_{\boldsymbol{u}_1}(\boldsymbol{X}_{\boldsymbol{u}_1}), \ldots, f_{\boldsymbol{u}_d}(\boldsymbol{X}_{\boldsymbol{u}_d})$ are d unspecified nonparametric functions, then it is said that $f(\boldsymbol{X})$ admits the *group additive structure* G. We also refer to (1) as a group additive model for $f(\boldsymbol{X})$. It is clear that the usual additive model is a special case with $G = \{(1), \ldots, (p)\}$.

Suppose X_{j_1} and X_{j_2} are two predictor variables. Intuitively, if X_{j_1} and X_{j_2} interact to each other, then they must appear in the same group in an reasonable group additive structure of $f(\boldsymbol{X})$. This implies that the group additive model could preserve the interactions between the predictor variables. On the other hand, if X_{j_1} and X_{j_2} belong to two different groups, then they do not interact with each other. In other words, the group additive model assumes no interaction between different groups. Therefore, in terms of accommodating interactions, the group additive model can be considered lying in the middle between the original additive model and the functional ANOVA or higher order interaction models. When the group sizes are small, for example all are less than or equal to 3, the group additive model can maintain the estimation efficiency and interpretability of the original additive model while avoiding the problem of a high order model discussed earlier.

However, in [17], there are two important issues not addressed. The first issue that the group additive structure may not be unique, which will lead to the nonidentifiability problem for the group additive model. (See discussion in Section 2.1). The second issue is that [17] has not proposed a systematic approach to identify the group additive structure. In this paper, we intend to resolve these two issues. To address the first issue, we rigorously define the *intrinsic group additive structure* for any square integrable function, which in some sense is the minimal group additive structure among all correct group additive structures for the function.

To address the second issue, we propose a general approach to simultaneously identifying the intrinsic group additive structure and estimating the nonparametric functions using kernel methods and Reproducing Kernel Hilbert Spaces (RKHSs). For a given group additive structure $G = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$, we first define the corresponding direct sum RKHS as $\mathcal{H}_G = \mathcal{H}_{\boldsymbol{u}_1} \oplus \cdots \oplus \mathcal{H}_{\boldsymbol{u}_d}$ where $\mathcal{H}_{\boldsymbol{u}_i}$ is the usual RKHS for the variables in $\boldsymbol{u}_j$ only for $j = 1, \ldots, d$. Based on the results on the complexity measure of RKHSs in the literature, we derive a tractable complexity measure of the direct sum RKHS $\mathcal{H}_G$ which is further used as the complexity measure of G. Then, the identification of the intrinsic group additive structure G and the estimation of the involved nonparametric functions can be performed through the following minimization problem:

$$\hat{f}, \hat{G} = \underset{f \in H_G, G}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (y_i - f(\boldsymbol{x}_i))^2 + \lambda \|f\|_{\mathcal{H}}^2 + \mu \boldsymbol{\mathcal{C}}(G).$$

We show that when the novel complexity measure of group additive structure $\boldsymbol{\mathcal{C}}(G)$ is used, the minimizer $\hat{G}$ is consistent for the intrinsic group additive structure as the size of the training sample

goes to infinity. We further develop two algorithms, one uses exhaustive search and the other employs a stepwise approach, for identifying true additive group structures under the small p and large p scenarios. Extensive simulation study and real data applications show that our proposed method can successfully recover the true additive group structures in a variety of model settings.

There exists a connection between our proposed group additive model and graphical models ([4], [10]). This is especially true when a sparse block structure is imposed [15]. However, a key difference exists. Let's consider the following example. $Y = \sin(X_1 + X_2^2 + X_3) + \cos(X_4 + X_5 + X_6^2) + \epsilon$. A graphical model typically considers the conditional dependence (CD) structure among all of the variables including $X_1, \ldots, X_6$ and Y, which is more complex than the group additive (GA) structure $\{(X_1, X_2, X_3), (X_4, X_5, X_6)\}$. The CD structure, once known, can be further examined to infer the GA structure. In this paper, we however proposed methods that directly target the GA structure instead of the more complex CD structure.

The rest of the paper is organized as follows. In Section 2, we rigorously formulate the problem of Group Additive Structure Identification (GASI) for nonparametric regression and propose the structural penalty method to solve the problem. In Section 3, we prove the selection consistency for the method. We report the experimental results based on simulation studies and real data applications in Section 4 and 5. Section 6 concludes this paper with discussion.

2 Method

2.1 Group Additive Structures

In the Introduction, we discussed that the group additive structure for $f(\boldsymbol{X})$ may not be unique. Here we give an example. Consider the following model $Y = 2 + 3X_1 + 1/(1 + X_2^2 + X_3^2) + \arcsin((X_4 + X_5)/2) + \epsilon$, where ϵ is the error independent of $\boldsymbol{X}$ with 0 mean. According to the definition, this model admits the group additive structure $G_0 = \{(1), (2, 3), (4, 5)\}$. Let $G_1 = \{(1, 2, 3), (4, 5)\}$ and $G_2 = \{(1, 4, 5), (2, 3)\}$. The model can also be said to admit G_1 and G_2. However, there exists a major difference between G_0, G_1 and G_2. While the groups in G_0 cannot be further divided into subgroups, both G_1 and G_2 contain groups that can be further split. We define the following partial order between group structures to characterize the difference.

Definition 1. *Let G and G' be two group additive structures. If for every group $\boldsymbol{u} \in G$ there is a group $\boldsymbol{v} \in G'$ such that $\boldsymbol{u} \subseteq \boldsymbol{v}$, then G is called a **sub group additive structure** of G'. This relation is denoted as $G \leq G'$. Equivalently, G' is a **super group additive structure** of G, denoted as $G' \geq G$.*

In the previous example, G_0 is a sub group additive structure of both G_1 and G_2. However, the order between G_1 and G_2 is not defined.

Let $\mathcal{X} := [0, 1]^p$ be the p-dimensional unit cube for all the predictor variables $\boldsymbol{X}$ and $P_{\boldsymbol{X}}$ be the probability distribution. For a group of predictor variables $\boldsymbol{u}$, we define the space of square integrable functions as $L^2_{\boldsymbol{u}}(\mathcal{X}) := \{g \in L^2_{P_{\boldsymbol{X}}}(\mathcal{X}) \mid g(\boldsymbol{X}) = f_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\}$, that is $L^2_{\boldsymbol{u}}$ contains the functions that only depend on the variables in group $\boldsymbol{u}$. Then the group additive model $f(\boldsymbol{X}) = \sum_{j=1}^{d} f_{\boldsymbol{u}_j}(\boldsymbol{X}_{\boldsymbol{u}_j})$ is a member of the direct sum function space defined as $L^2_G(\mathcal{X}) := \oplus_{\boldsymbol{u} \in G} L^2_{\boldsymbol{u}}(\mathcal{X})$. Let $|\boldsymbol{u}|$ be the cardinality of the group $\boldsymbol{u}$. If $\boldsymbol{u}$ is the only group in a group additive structure and $|\boldsymbol{u}| = p$, then $L^2_{\boldsymbol{u}} = L^2_G$ and f is a fully non-parametric function.

The following proposition shows that the order of two different group additive structures is preserved by their corresponding square integrable function spaces.

Proposition 1. *Let G_1 and G_2 be two group additive structures. If $G_1 \leq G_2$, then $L^2_{G_1} \subseteq L^2_{G_2}$. Furthermore, if $X_1, \ldots, X_p$ are independent and $G_1 \neq G_2$, then $L^2_{G_1} \subset L^2_{G_2}$.*

Definition 2. *Let $f(\boldsymbol{X})$ be an square integrable function. For a group additive structure G, if there is a function $f_G \in L^2_G$ such that $f_G = f$, then G is called an **amiable group additive structure** for f.*

In the example discussed in the beginning of the subsection, G_0, G_1 and G_2 are all amiable group structures. So amiable group structures may not be unique.

Proposition 2. *Suppose G is an amiable group additive structure for f. If there is a second group additive structure G' such that $G \leq G'$, then G' is also amiable for f.*

We denote the collection of all amiable group structures for $f(\boldsymbol{X})$ as $\mathcal{G}^a$, which is partially ordered and complete. Therefore, there exists a minimal group additive structure in $\mathcal{G}^a$, which is the most concise group additive structure for the target function. We state this result as a theorem.

Theorem 1. *Let $\mathcal{G}^a$ be the set of amiable group additive structures for f. There is a unique minimal group additive structure $G^* \in \mathcal{G}^a$ such that $G^* \leq G$ for all $G \in \mathcal{G}^a$, where the order is given by Definition 1. G^* is called the* ***intrinsic group additive structure*** *for f.*

For statistical modeling, G^* achieves the greatest dimension reduction for the relationship between Y and $\boldsymbol{X}$. It induces the smallest function space which includes the model. In the previous example, we have G_0 being the intrinsic group additive structure. If $G^* = G_0$ is known, one only needs to estimate one univariate and two bivariate non-parametric functions. Although G_1 and G_2 are both amiable, they both require fitting a three-dimensional non-parametric functions. This is both computationally and statistically inefficient. In general, the intrinsic group structure can help much mitigate the curse of dimensionality while improving both efficiency and interpretability of high dimensional nonparametric regression.

2.2 Kernel Method with Known Intrinsic Group Additive Structure

Consider $f(\boldsymbol{X}) = \mathbb{E}[Y|\boldsymbol{X}]$. Suppose the intrinsic group additive structure for $f(\boldsymbol{X})$ is known to be $G^* = \{\boldsymbol{u}_j\}_{j=1}^d$, that is, $f(\boldsymbol{X}) = f_{\boldsymbol{u}_1}(\boldsymbol{X}_{\boldsymbol{u}_1}) + \cdots + f_{\boldsymbol{u}_d}(\boldsymbol{X}_{\boldsymbol{u}_d})$. Therefore, estimating f is essentially to estimate the functions $f_{\boldsymbol{u}_1}, f_{\boldsymbol{u}_2}, \ldots, f_{\boldsymbol{u}_d}$. We will use the kernel method. Let $(K_{\boldsymbol{u}_j}, \mathcal{H}_{\boldsymbol{u}_j})$ be the kernel and its corresponding RKHS for the j-th group $\boldsymbol{u}_j$. Then using kernel methods is to solve

$$\hat{f}_{\lambda,G^*} = \underset{f_{G^*} \in \mathcal{H}_{G^*}}{\arg\min} \left\{ \frac{1}{n} \left(y_i - f_{G^*}(\boldsymbol{x}_i) \right)^2 + \lambda \| f_{G^*} \|^2_{\mathcal{H}_{G^*}} \right\}, \tag{2}$$

where $\mathcal{H}_{G^*} := \{ f = \sum_{j=1}^d f_{\boldsymbol{u}_j} \mid f_{\boldsymbol{u}_j} \in \mathcal{H}_{\boldsymbol{u}_j} \}$. The subscripts are used on RHS to explicitly indicate the dependence of the soluction on the group additive structure G^* and tuning parameter λ. The solution is searched in the corresponding direct sum RKHS and can be solved by KRR.

In general, an RKHS is usually smaller than the L^2 space defined on the same input domain. So, it is not always that $\hat{f}_{\lambda,G^*} \equiv f$, and in fact a bias exists. However, one can choose to use kernels $K_{\boldsymbol{u}_j}$ that are universal in the sense that their corresponding RKHSs are dense in the L^2 spaces (see [22], [5]). Two examples of universal kernel are Gaussian and Laplace. By using universal kernels, not only can the bias of $\hat{f}_{\lambda,G^*}$ reduces to zero as n goes to infinity, but also can $\hat{f}_{\lambda,G^*}$ recover the structural properties such as the group additive structure of $f(\boldsymbol{X})$. This is the fundamental reason for the consistency property of our proposed method to identify the intrinsic group additive structure.

2.3 Identification of Unknown Intrinsic Group Additive Structure

2.3.1 Penalization on Group Additive Structures

The success of the kernel method discussed in the previous subsection hinges on the knowledge of the intrinsic group additive structure G^*. In practice, however, G^* is seldom known, and it may be of primary interest to identify G^* while fitting a nonparametric regression for $\mathbb{E}[Y|\boldsymbol{X}]$ as discussed earlier. Recall that in Subsection 2.1, we have shown that G^* exists and unique. The other possible group additive structures belong to two categories, amiable and non-amiable.

Let's consider an arbitrary non-amiable group additive structure $G \in \mathcal{G} \setminus \mathcal{G}^a$ first. Suppose G is used in the place of G^* in the kernel method (2). The resulting fitted function $\hat{f}_{\lambda,G}$, as an estimator of f, will have a systematic bias because the L^2 distance between any function f_G in $\mathcal{H}_G$ and the true function f will be bounded below. The bias remains regardless of the size of the training sample. In other words, using a non-amiable group additive structure will result in poor fitting of the model.

Next we consider an arbitrary amiable group additive structure G^a. Suppose G^a is used in the place of G^* in (2). Recall that because G^a is amiable, therefore for the true function $f(\boldsymbol{X})$, we have $f_{G^*} = f_{G^a}$ almost surely. The bias of the resulting fitted function $\hat{f}_{\lambda,G^a}$ will vanish as the sample size increases. Although their asymptotic rates are in general different, under fixed sample size n, simply using goodness of fit will not be able to distinguish G^a from G^*. The key difference between G^* and G^a is their structural complexities, that is, G^* is the smallest among all amiable

structures (i.e. $G^* \leq G, \forall G \in \mathcal{G}^a$). Suppose a proper measure of the complexity of a group additive structure G can be defined (to be addressed in the next section) and is denoted as $\boldsymbol{\mathcal{C}}(G)$. We can then incorporate $\boldsymbol{\mathcal{C}}(G)$ into (2) as an additional penalty term and change the kernel method to the following structure-penalized method.

$$\hat{f}_{\lambda,\mu}, \hat{G} = \underset{f_G \in \mathcal{H}_G, G}{\arg\min} \left\{ \frac{1}{n} \left(y_i - f_G(\boldsymbol{x}_i)\right)^2 + \lambda \|f_G\|^2_{\mathcal{H}_G} + \mu \boldsymbol{\mathcal{C}}(G) \right\}, \tag{3}$$

where $\mathcal{H}_G := \{f = \sum_{j=1}^{d} f_{\boldsymbol{u}_j} \mid f_{\boldsymbol{u}_j} \in \mathcal{H}_{\boldsymbol{u}_j}\}$. It is clear that the only difference between (2) and (3) is the term $\mu\boldsymbol{\mathcal{C}}(G)$. As discussed above, the intrinsic group additive structure G^* can achieve the goodness of fit represented by the first two terms in (3) and the penalty on the structural complexity represented by the last term. Therefore, as the sample size n increases, by properly choosing the tuning parameters, we expect that $\hat{G}$ is consistent in that the probability $\hat{G} = G^*$ increases to one as n increases (see the Theory Section below). We refer to (3) as the structure-penalized kernel method. In the next section, we derive a tractable complexity measure for a group additive structure.

2.3.2 Complexity Measure of Group additive Structure

It is tempting to propose an intuitive complexity measure for a group additive structure $\boldsymbol{\mathcal{C}}(\cdot)$ such that $\boldsymbol{\mathcal{C}}(G_1) \leq \boldsymbol{\mathcal{C}}(G_2)$ whenever $G_1 \leq G_2$. The intuition however breaks down or at least becomes less clear when the order between G_1 and G_2 cannot be defined. From Proposition 1, it is known that when $G_1 < G_2$, we have $L^2_{G_1} \subset L^2_{G_2}$. It is not difficult to show that it is also true that when $G_1 < G_2$, then $\mathcal{H}_{K,G_1} \subset \mathcal{H}_{K,G_2}$. This observation motivates us to define the structural complexity measure of G through the measure of the capacity of its corresponding RKHS $\mathcal{H}_G$.

There exist a number of different types of complexity measures for RKHSs in the literature, including entropy [25], VC dimensions [24], Rademacher complexity [3], and covering numbers ([21], [25]). After investigating and comparing these different measures, we use covering number to design a practically convenient complexity measure for group additive structures.

It is known that an RKHS $\mathcal{H}_K$ can be embedded in the continuous function space $\mathcal{C}(\mathcal{X})$ (see [18], [26]), with the inclusion mapping denoted as $I_K : \mathcal{H}_K \to \mathcal{C}(\mathcal{X})$. Let $\mathcal{H}_{K,r} = \{h : \|h\|_{H_k} \leq r, \text{ and } h \in \mathcal{H}_K\}$ be an r-ball in $\mathcal{H}_K$ and $\overline{I\left(\mathcal{H}_{K,r}\right)}$ be the closure of $I\left(\mathcal{H}_{K,r}\right)$ in $\mathcal{C}(\mathcal{X})$. One way to characterize the capacity or complexity of $\mathcal{H}_K$ is through the covering number of $\overline{I\left(\mathcal{H}_{K,r}\right)}$ in $\mathcal{C}(\mathcal{X})$, denoted as $\mathcal{N}(\epsilon, \overline{I\left(\mathcal{H}_{K,r}\right)}, d_\infty)$, which is the smallest cardinalty of a cover or subset S of $\mathcal{C}(\mathcal{X})$ such that $\overline{I\left(\mathcal{H}_{K,r}\right)} \subset \cup_{s\in S}\{t \in \mathcal{C}(\mathcal{X}) : d_\infty(t,s) \leq \epsilon\}$. Here ϵ is any small positive value, and d_∞ is the usual sup norm of $\mathcal{C}(\mathcal{X})$.

The exact formula for $\mathcal{N}(\epsilon, \overline{I\left(\mathcal{H}_{K,r}\right)}, d_\infty)$ is in general not available. Under certain conditions, upper bounds for $\mathcal{N}(\epsilon, \overline{I\left(\mathcal{H}_{K,r}\right)}, d_\infty)$ have been obtained in the literature. One such upper bound is presented as follows.

When K is a convolution kernel, i.e. $K(x,t) = k(x-t)$, and the Fourier transform of k decays exponentially, then, it is given in [26] that

$$\ln \mathcal{N}\left(\epsilon, \overline{I(\mathcal{H}_{K,r})}, d_\infty\right) \leq C_{k,p} \left(\ln \frac{r}{\epsilon}\right)^{p+1} \tag{4}$$

where $C_{k,p}$ is a constant depending on the kernel function k and input dimension p. In particular, when K is a Gaussian kernel, [25] has obtained more elaborate upper bounds.

The upper bound in (4) depends on r and ϵ through $\ln(r/\epsilon)$. When $\epsilon \to 0$ with r fixed (e,g. $r = 1$ when a unit ball is considered), $(\ln(r/\epsilon))^{p+1}$ becomes the dominant factor in the upper bound. According to [11], the growth rate of $\mathcal{N}\left(\epsilon, I_K\right)$ or its logarithmic version can be viewed as a measure of the complexity of RKHS. So we use $(\ln(r/\epsilon))^{p+1}$ as the complexity measure, which is equivalent to α^{p+1} where α is the reparameterization of $\ln(r/\epsilon)$. Let $\mathcal{C}(\mathcal{H}_k)$ denote the complexity measure of $\mathcal{H}_k$, which is defined as $\boldsymbol{\mathcal{C}}(\mathcal{H}_k) = (\ln(r/\epsilon))^{p+1} = \alpha(\epsilon)^{p+1}$. We know ϵ is the radius of a covering ball, which is the unit of measurement we use to quantify the complexity. The complexity of two RKHSs with different input dimensions are easier to be differentiated when ϵ is small. This gives an interpretation of α.

We have defined a complexity measure for a general RKHS. In Problem (3), the model space $\mathcal{H}_G$ is a direct sum of a number of RKHSs. Let $G = \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d\}$; let $\mathcal{H}_G, \mathcal{H}_{\boldsymbol{u}_1}, \ldots, \mathcal{H}_{\boldsymbol{u}_d}$ be the RKHSs

corresponding to $G, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$, respectively; let $I_G, I_{\boldsymbol{u}_1}, \ldots, I_{\boldsymbol{u}_d}$ be the inclusion mappings of $\mathcal{H}_G, \mathcal{H}_{\boldsymbol{u}_1}, \ldots, \mathcal{H}_{\boldsymbol{u}_d}$ into $\mathcal{C}(\mathcal{X})$. Then, we have the following proposition.

Proposition 3. *Let G be a group additive structure and $\mathcal{H}_G$ be the induced direct sum RKHS defined in (3). Then, we have the following inequality relating the covering number of $\mathcal{H}_G$ and the covering numbers of $\mathcal{H}_{\boldsymbol{u}_j}$*

$$\ln \mathcal{N}(\epsilon, I_G, d_\infty) \leq \sum_{j=1}^{d} \ln \mathcal{N}\left(\frac{\epsilon}{|G|}, I_{\boldsymbol{u}_j}, d_\infty\right), \tag{5}$$

where $|G|$ denotes the number of groups in G.

By applying Proposition 3 and the parameterized upper bound, we have $\ln \mathcal{N}(\epsilon, I_G, d_\infty) = \mathcal{O}\left(\sum_{\boldsymbol{u} \in G} \alpha(\epsilon)^{|\boldsymbol{u}|+1}\right)$, where we explicitly indicate the dependency of α on ϵ. Now we could use the rate as the explicit expression of the complexity measure $\boldsymbol{\mathcal{C}}(G)$ in Problem (3), that is $\boldsymbol{\mathcal{C}}(G) = \sum_{j=1}^{d} \alpha(\epsilon)^{|\boldsymbol{u}_j|+1}$. Recall that there is another tuning parameter μ which controls the effect of the complexity of group structure has on the penalized risk. By factoring out the common 1 in the exponent for all groups and combining it with μ, we could further simplify the penalty's expression. Thus, we have the following explicit formulation for AGSI which simultaneously solves the non-parametric regression problem.

$$\hat{f}_{\lambda,\mu}, \hat{G} = \underset{f_G \in \mathcal{H}_G, G}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - f_G(\boldsymbol{x}_i)\right)^2 + \lambda \|f_G\|_{\mathcal{H}_G}^2 + \mu \sum_{j=1}^{d} \alpha^{|\boldsymbol{u}_j|} \right\}. \tag{6}$$

2.4 Estimation

We assume that the value of λ is pre-specified. In practice, this parameter can be tuned separately. If the values of μ and α are given, Problem (6) can be solved by following a two-step procedure. First, when the group structure G is given, the functions $f_{\boldsymbol{u}}$ can be estimated by using KRR to solve the following problem

$$\hat{\mathcal{R}}_G^\lambda = \min_{f_G \in \mathcal{H}_G} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(y_i - f_G(\boldsymbol{x}_i)\right)^2 + \lambda \left\|f_G\right\|_{\mathcal{H}_G}^2 \right\}. \tag{7}$$

Second, the optimal group structure is chosen to achieve both small fitting error and the complexity penalty, i.e.

$$\hat{G} = \underset{G \in \mathcal{G}}{\arg\min} \left\{ \hat{\mathcal{R}}_G^\lambda + \mu \sum_{j=1}^{d} \alpha^{|\boldsymbol{u}_j|} \right\}. \tag{8}$$

The two-step procedure above is expected to identify the intrinsic group structure, that is, $\hat{G} = G^*$. Recall a group structure belongs to one of the three categories, intrinsic, amiable, or non-amiable structures. If G is non-amiable, then $\hat{\mathcal{R}}_G^\lambda$ is expected to be large, because G is a wrong structure and will result in a biased estimate. If G is amiable, though $\hat{\mathcal{R}}_G^\lambda$ is expected to be small, the complexity penalty of G is larger than that for G^*. As a consequence, only G^* can simultaneously achieve a small $\hat{\mathcal{R}}_{G^*}^\lambda$ and a relatively small complexity penalty. Therefore, when the sample size is large enough, we expect $\hat{G} = G^*$ with high probability. If the values of the turning parameters μ and α are not given, a separate validation set can be used to select tuning parameters. The estimation is summarized in Algorithm 1

Algorithm 1: Exhaustive Search w/ Validation

Split data into training ($\mathcal{T}$) and validation ($\mathcal{V}$) sets.
for (μ, α) in grid **do**
  for $G \in \mathcal{G}$ **do**
    $\hat{\mathcal{R}}_G, \hat{f}_G \leftarrow$ solve (7) using G;
    Calculate the sum in (8), denoted by $\hat{\mathcal{R}}_G^{\text{pen},\mu,\alpha}$;
  end for
  $\hat{G}^{\mu,\alpha} \leftarrow \arg\min_{G\in\mathcal{G}} \hat{\mathcal{R}}_G^{\text{pen},\mu,\alpha}$;
  $\hat{y}^{\mathcal{V}} \leftarrow \hat{f}_{\hat{G}^{\mu,\alpha}}(\boldsymbol{x}^{\mathcal{V}})$;
  $e^2_{\hat{G}^{\mu,\alpha}} \leftarrow \|y^{\mathcal{V}} - \hat{y}^{\mathcal{V}}\|^2$;
end for
$\mu^*, \alpha^* \leftarrow \arg\min_{\mu,\alpha} e^2_{\hat{G}^{\mu,\alpha}}$;
$G^* \leftarrow \hat{G}^{\mu^*,\alpha^*}$;

Algorithm 2: Basic Backward Stepwise

State with the group structure $\{(1, \ldots, p)\}$;
Solve (6) and obtain its minimum value $\hat{\mathcal{R}}_G^{\text{pen}}$;
for each predictor variable j **do**
  $G' \leftarrow$ either split j as a new group or add to an existing group;
  Solve (6) and obtain its minimum value $\hat{\mathcal{R}}_{G'}^{\text{pen}}$;
  if $\hat{\mathcal{R}}_{G'}^{\text{pen}} < \hat{\mathcal{R}}_G^{\text{pen}}$ **then**
    Keep G' as the new group structure;
  end if
end for
return G';

Algorithm 1 selects the group additive structure by compare the results of all possible group structures. When a model contains a large number of predictor variables, such exhaustive search suffers high computational cost. In order to apply GASI on a large model, we propose a backward stepwise algorithm which is illustrated in Algorithm 2.

3 Theory

In this section, we prove that the estimated group additive structure $\hat{G}$ as a solution to (6) is consistent, that is the probability $P(\hat{G} = G^*)$ goes to 1 as the sample size n goes to infinity. As we discussed before, when a non-amiable group additive structure is used, the solution of a usual kernel nonparametric regression problem has a non-zero bias. While all amiable group additive structures give unbiased estimates, using the intrinsic group additive structure will enjoy the fastest rate of convergence. Thus, the new complexity penalty is used to filter out all amiable group structures with slow convergence rate. We provide the main theorems in this section. The proof and supporting lemmas are included in the supplemental document.

Let $\mathcal{R}(f_G) := \mathbb{E}[(Y - f(\boldsymbol{X}))^2]$ denote the population risk of function $f \in \mathcal{H}_G$, and $\hat{\mathcal{R}}(f) := \frac{1}{n}\sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2$ be the empirical risk. First, we show that for any given amiable group additive structure $G \in \mathcal{G}^a$, its optimized empirical risk $\hat{\mathcal{R}}(\hat{f}_G)$ converges in probability to the optimal population risk $\mathcal{R}(f^*_{G^*})$ achieved by the intrinsic group additive structure. Here $\hat{f}_G$ denotes the minimizer of the regularized empirical risk (7) when the group additive structure G is used, and $f^*_{G^*}$ denotes the minimizer of the population risk when the intrinsic group structure is used. The result is given below as Proposition 4.

Proposition 4. *Let G^* be the intrinsic group additive structure, $G \in \mathcal{G}^a$ a given amiable group structure, and $\mathcal{H}_{G^*}$ and $\mathcal{H}_G$ the respective direct sum RKHSs. If $\hat{f}_G^\lambda \in \mathcal{H}_G$ is the optimal solution of Problem (7), then for any $\epsilon > 0$, we have*

$$P\left(|\widehat{\mathcal{R}}(\hat{f}_G) - \mathcal{R}(f^*_{G^*})| > \epsilon\right) \le 12n \cdot \exp\left\{\sum_{\boldsymbol{u}\in G} \ln\mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, d_\infty\right) - \frac{\epsilon^2 n}{144}\right\} +$$
$$12n \cdot \exp\left\{\sum_{\boldsymbol{u}\in G} \ln\mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, d_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\}. \quad (9)$$

Note that λ_n in (21) must be chosen such that $\frac{\epsilon}{24} - \frac{\lambda_n\|f^*_{G^*}\|^2}{12}$ is positive. For any given ϵ, when n is sufficiently large, the exponents of the two terms in (21) will become negative. When n further increases, both of the terms in (21) will decrease exponentially to zero. Therefore, Proposition 4 implies that $\widehat{\mathcal{R}}(\hat{f}_G)$ converges to $\mathcal{R}(f^*_{G^*})$ in probability. For a fixed p and intrinsic group additive structure, the number of amiable group additive structures is finite. Using a Bonferroni type of

ID	Model	Intrinsic Group Structure
M1	$y = 2x_1 + x_2^2 + x_3^3 + \sin(\pi x_4) + \log(x_5 + 5) + \lvert x_6\rvert + \epsilon$	$\{(1), (2), (3), (4), (5), (6)\}$
M2	$y = \frac{1}{1+x_1^2} + \arcsin\left(\frac{x_2+x_3}{2}\right) + \arctan\left((x_4 + x_5 + x_6)^3\right) + \epsilon$	$\{(1), (2,3), (4,5,6)\}$
M3	$y = \arcsin\left(\frac{x_1+x_3}{2}\right) + \frac{1}{1+x_2^2} + \arctan\left((x_4 + x_5 + x_6)^3\right) + \epsilon$	$\{(1,3), (2), (4,5,6)\}$
M4	$y = x_1 \cdot x_2 + \sin((x_3 + x_4) \cdot \pi) + \log(x_5 \cdot x_6 + 10) + \epsilon$	$\{(1,2), (3,4), (5,6)\}$
M5	$y = \exp\left\{\sqrt{x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 + x_6^2}\right\} + \epsilon$	$\{(1,2,3,4,5,6)\}$

Table 1: Selected models for the simulation study using the exhaustive search method and the corresponding additive group structures.

technique, we can in fact obtain a uniform upper bound for all of the amiable group additive structures in $\mathcal{G}^a$. This result is stated in the following theorem.

Theorem 2. *Let $\mathcal{G}^a$ be the set of all amiable group structures. For any $\epsilon > 0$ and $n > 2/\epsilon^2$, we have*

$$P\left(\sup_{G\in\mathcal{G}^a} |\widehat{\mathcal{R}}_g(\hat{f}_G^{\lambda}) - \mathcal{R}_g(f_{G^*}^*)| > \epsilon\right) \leq 12n|\mathcal{G}^a| \cdot \left[\exp\left\{\max_{G\in\mathcal{G}^a} \ln\mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, d_\infty\right) - \frac{\epsilon^2 n}{144}\right\}\right.$$
$$\left. + \exp\left\{\max_{G\in\mathcal{G}^a} \ln\mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, d_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f_{G^*}^*\|^2}{12}\right)^2\right\}\right] \tag{10}$$

Theorem 2 implies that the convergence of $\widehat{\mathcal{R}}(\hat{f}_G)$ to $\mathcal{R}(f_{G^*}^*)$ in probability is uniform for G in $\mathcal{G}^a$.

Next we consider a non-amiable group additive structure $G' \in \mathcal{G} \setminus \mathcal{G}^a$. It turns out that $\widehat{\mathcal{R}}(\hat{f}_G)$ fails to converge to $\mathcal{R}(f_{G^*}^*)$ in probability, and $|\widehat{\mathcal{R}}(\hat{f}_G) - \mathcal{R}(f_{G^*}^*)|$ converges to a positive constant in probability. Furthermore, because the number of non-amiable group additive structures is finite, we can show that $|\widehat{\mathcal{R}}(\hat{f}_G) - \mathcal{R}(f_{G^*}^*)|$ is uniformly bounded below from zero with probability going to 1. We state the results as the next theorem.

Theorem 3. *(i) For a non-amiable group structure $G' \in \mathcal{G} \setminus \mathcal{G}^a$, there exists a constant $C > 0$ such that $|\widehat{\mathcal{R}}_g(\hat{f}_{G'}^{\lambda}) - \mathcal{R}_g(f_{G^*}^*)|$ converges to C in probability.*
(ii) There exits a constant $\tilde{C}$ such that $P(|\widehat{\mathcal{R}}_g(\hat{f}_{G'}^{\lambda}) - \mathcal{R}_g(f_{G^*}^*)| > \tilde{C}$ for all $G' \in \mathcal{G} \setminus \mathcal{G}^a)$ goes to 1 as n goes to infinity.

By combining Theorem 2 and Theorem 3, it is not difficult to show that the probability that the solution of (6) $\hat{G}$ is not equal to the intrinsic group additive structure goes to zero as n goes to infinity. The structural penalty helps to distinguish amiable structures from the intrinsic group additive structure. We state this result in the following theorem.

Theorem 4. *Let $\lambda_n * n \to 0$. By choosing a proper tuning parameter $\mu > 0$ for the structural penalty , the estimated group structure $\hat{G}$ is consistent for the intrinsic group additive structure G^*, that is, $P(\hat{G} = G^*)$ goes to one as the sample size n goes to infinity.*

4 Simulation

In this section, we evaluate the performance of GASI using synthetic data. Table 1 shows the five models we are using. Observations of $\boldsymbol{X}$ are simulated independently from $N(0,1)$ in M1, Unif$(-1,1)$ in M2 and M3, and Unif$(0,2)$ in M4 and M5. The noise ϵ is i.i.d. $N(0, 0.01^2)$. The grid values of μ are equally spaced in $[1\text{e}{-10}, 1/64]$ on a log-scale and each α is an integer in $[1, 10]$.

We first show that GASI has the ability to identify the intrinsic additive group structure. For this, we apply the two-step procedure for each (μ, α) pair multiple times. If there are (μ, α) pairs for each model that its true group structure can be often identified, then GASI has the power to identify true group structures. We also apply Algorithm 1 which uses an additional validation set to select the parameters. The simulation is repeated 100 times for each model.The frequency of the true group structure being identified is calculated for each (μ, α).

Model	Max freq.	μ	α	Max freq.	μ	α	Max freq.	μ	α
M1	100	1.2500e-06	10	59	1.2500e-06	4	99	1.5625e-02	10
M2	97	1.2500e-06	8	89	1.2500e-06	7	70	1.3975e-04	9
M3	97	1.2500e-06	9	89	1.2500e-06	7	65	1.3975e-04	8
M4	100	1.2500e-06	7	99	1.2500e-06	4	1	1.3975e-04	8
M5	100	1.2500e-06	1	100	1.2500e-06	1	100	1.2500e-06	1

Table 2: Maximum frequencies that the intrinsic group additive structures are identified for the five models using exhaustive search algorithm without parameter tuning (left panel), with parameter tuning (middle panel) and stepwise algorithm (right panel). If different pairs share the same maximum frequency, a pair is randomly chosen.

In Table 2, we report the maximum frequency and the corresponding (μ, α) for each model. The complete results are included in the supplemental document. It can be seen from the left frequency panel that the intrinsic group additive structures can be successfully identified. When the parameters are tuned, the middle panel shows that the performance of Model 1 deteriorated. This might be caused by the estimation method (Kernel Ridge Regression to solve Problem (7)) used in the algorithm. It could also be affected by λ.

When the number of predictor variables increases, we propose to use a backward stepwise algorithm. We apply Algorithm 2 on the same models. The results are reported in the right panel in Figure 2. The true group structures could be identified most of time for Model 1, 2, 3, 5. The result of Model 4 is not satisfying. Since stepwise algorithm is greedy, it is possible that the true group structures were never visited. Further research is needed to better understand the role of the complexity penalty in stepwise algorithms.

5 Real Data

In this section, we apply GASI on two real data sets, which are both available in the UCI repository.

The first data set is the Boston Housing data. It includes 13 predictor variables which are used to predict the house median value. The sample size is 506. Our goal is to identify a probable group structure for the predictor variables. The backward algorithm is used and the tuning parameters μ and α are selected via 10-fold CV. The group additive structure that achieves the lowest average validation error is $\{(1, 6), (2, 11), (3), (4, 9), (5, 8), (7, 13), (10, 12)\}$, which is used for further investigation. Then the nonparametric functions for each group were estimated using the whole data set. Because each group contains no more than two variables, the estimated functions can be visualized. Figure 1 shows the selected results.

It is interesting to see some patterns emerging in the plots. The top-left plot shows the function of the average number of rooms per dwelling and per capita crime rate by town. We can see the house value increases with more rooms and decreases as the crime rate increases. However, when the crime rate is low, smaller sized houses (4 or 5 rooms) seem to be preferred. The top-right plot shows that there is a changing point in terms of how house value is related to the size of non-retail business in the area. The value initially drops when the percentage of non-retail business is small, then increases at around 8%. The increase in the value might be due to the high demand of housing from the employees of those business.

The second data set is the communities and crime data (unnormalized). It combines socio-economic, law enforcement, and crime data collected by US government agencies. There are 2215 samples and 147 variables with missing values. We choose Number of Murders in 1995 to be the response in this study and investigate its relationship between the predictor variables. We removed the observations with missing values.

To deal with the large number of predictor variables, a screening procedure is used to select the most related variables. We fit OKGT for each of the 122 predictor against the response, then keep the variables with $R^2 > 0.99$. This ensures that the selected predictors are highly dependent to the response. The screening procedure selects 23 predictor variables. We also remove the samples with

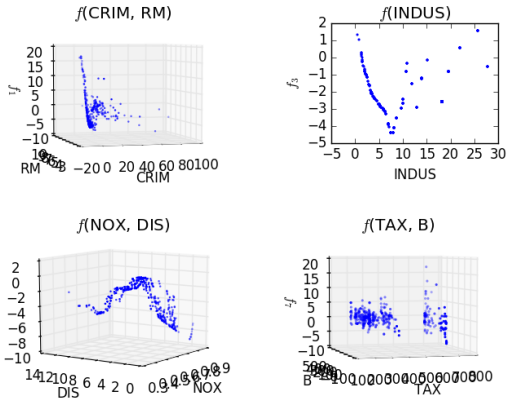

Figure 1: Estimated transformation functions for selected groups. Top-left: group $(1, 6)$, top-right: group (3), bottom-left: group $(5, 8)$, bottom-right: group $(10, 12)$.

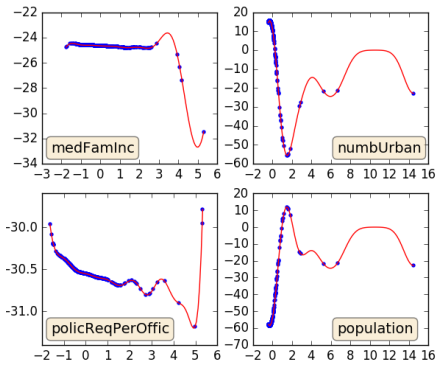

Figure 2: Selected results for the communities and crime data where the number of murders is the response. The blue dots are the transformed observation of the predictor variable. The red line is the estimated function.

missing values, which reduced sample size to 343. Then the backward algorithm is applied on the prepared data set.

The procedure selected the fully additive group structure. Figure 7 includes selected results which show highly nonlinear effect of each predictor variable. The first plot shows that the effect of Median Family Income is almost zero until it reaches the high end where murders drop dramatically. The second plot shows an interesting pattern for Total Requests for Police per Police Officer. As the number of requests increases, the number of murders initially decreases slowly. One reason for this is that increasing requests cause more presence of police in the area which is helpful to control crimes. However, murders increase quickly as the number of requests enters the high range. An explanation for this is that the surging number of requests for police is due to the low security and high murder rate in the area.

6 Discussion

We use group additive model for nonparametric regression and propose a RKHS complexity penalty based approach for identifying the intrinsic group additive structure. There are two main directions for future research. First, our penalty is based on the covering number of RKHSs. It is of interest to know if there exist other more effective penalty. Second, the current backward stepwise algorithm may become unstable and fail to achieve the potential in identifying the true additive group structure. It is of great interest to further improve the proposed method that can be applied in general high dimensional nonparametric regression.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[2] F. R. Bach. Exploring large feature spaces with hierarchical multiple kernel learning. In *Advances in neural information processing systems*, pages 105–112, 2009.

[3] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *The Journal of Machine Learning Research*, 3:463–482, 2003.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] C. Carmeli, E. De Vito, A. Toigo, and V. Umanitá. Vector valued reproducing kernel hilbert spaces and universality. *Analysis and Applications*, 8(01):19–61, 2010.

[6] F. Cucker and S. Smale. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2001.

[7] C. Gu. *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media, 2013.

[8] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical science*, pages 297–310, 1986.

[9] K. Kandasamy and Y. Yu. Additive approximations in high dimensional nonparametric regression via the salsa. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, ICML'16, pages 69–78. JMLR.org, 2016.

[10] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. The MIT Press, 2009.

[11] T. Kühn. Covering numbers of Gaussian reproducing kernel Hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

[12] T. Kühn. Covering numbers of gaussian reproducing kernel hilbert spaces. *Journal of Complexity*, 27(5):489–499, 2011.

[13] F. Kuo, I. Sloan, G. Wasilkowski, and H. Woźniakowski. On decompositions of multivariate functions. *Mathematics of computation*, 79(270):953–966, 2010.

[14] Y. Lin, H. H. Zhang, et al. Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics*, 34(5):2272–2297, 2006.

[15] B. M. Marlin and K. P. Murphy. Sparse gaussian graphical models with unknown block structure. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 705–712, New York, NY, USA, 2009. ACM.

[16] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[17] C. Pan, Q. Huang, and M. Zhu. Optimal kernel group transformation for exploratory regression analysis and graphics. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 905–914. ACM, 2015.

[18] T. Poggio and C. Shelton. On the mathematical foundations of learning. *American Mathematical Society*, 39(1):1–49, 2002.

[19] J. O. Ramsay and B. W. Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Citeseer, 2002.

[20] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[21] A. J. Smola and B. Schölkopf. *Learning with kernels*. Citeseer, 1998.

[22] I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[23] C. J. Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

[24] V. Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.

[25] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739 – 767, 2002.

[26] D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Appendix

A. Theorem, Proposition, Lemma and Proof

Proposition 1. *Let G_1 and G_2 be two group additive structures. If $G_1 \leq G_2$, then $L^2_{G_1} \subseteq L^2_{G_2}$. Furthermore, if $X_1, \ldots, X_p$ are independent and $G_1 \neq G_2$, then $L^2_{G_1} \subset L^2_{G_2}$.*

Proof. First we prove the first part.

Since $f \in L^2_{G_1}$, we have $f = \sum_{\boldsymbol{u} \in G_1} f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})$.

If $G_1 \cap G_2 \neq \emptyset$, then for each $\boldsymbol{u} \in G_1 \cap G_2$, it is true that $f_{\boldsymbol{u}} \in L^2_{G_2}$.

If $\boldsymbol{u} \notin G_1 \cap G_2$ and $\boldsymbol{u} \in G_1 \setminus G_2$, because $G_1 \leq G_2$, there exists $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_k \in G_2 \setminus G_1$ for some $k < |G_2|$ such that $\boldsymbol{v} := \boldsymbol{u} \cup \boldsymbol{u}_1 \cup \cdots \cup \boldsymbol{u}_k \in G_2$. Since

$$L^2([0,1]^{|\boldsymbol{u}|}) \oplus L^2([0,1]^{|\boldsymbol{u}_1|}) \oplus \cdots \oplus L^2([0,1]^{|\boldsymbol{u}_k|}) \subseteq L^2([0,1]^{|\boldsymbol{v}|}), \tag{11}$$

by induction, we have the desired result.

The sub-additivity in (11) is true because for two groups $\boldsymbol{u}$ and $\boldsymbol{v}$ in a group structure G, we have

$$\begin{aligned}
&\int (f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) + f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}))^2 p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \\
=&\int f^2_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} + \int f^2_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \\
&\qquad + 2 \int f_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) f_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \\
\leq&\int f^2_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} + \int f^2_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \\
&\qquad + 2 \left(\int f^2_{\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \right)^{1/2} \cdot \left(\int f^2_{\boldsymbol{v}}(\boldsymbol{x}_{\boldsymbol{v}}) p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \right)^{1/2} \\
<&\infty
\end{aligned}$$

The second to the last inequality is due to Holder's inequality with $p = q = 2$.

We further need to show the proper part (i.e. strict subset).

For $\boldsymbol{u}, \boldsymbol{v} \in G_1$, $\boldsymbol{u}, \boldsymbol{v} \notin G_2$, $\boldsymbol{u} \cup \boldsymbol{v} \in G_2$, we need to show that there is a function $h(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) \in L^2([0,1]^{|\boldsymbol{u} \cup \boldsymbol{v}|})$ which does not belong to $L^2_{\boldsymbol{u}} \oplus L^2_{\boldsymbol{v}}$. That is

$$\inf_{\substack{f \in L^2_{\boldsymbol{u}} \\ g \in L^2_{\boldsymbol{v}}}} \int \left(h(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) - f(\boldsymbol{x}_{\boldsymbol{u}}) - g(\boldsymbol{x}_{\boldsymbol{v}}) \right)^2 p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} > 0 \tag{12}$$

Define the following functional of f and g as

$$F(f, g) := \int \left(h(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) - f(\boldsymbol{x}_{\boldsymbol{u}}) - g(\boldsymbol{x}_{\boldsymbol{v}}) \right)^2 p(\boldsymbol{x}_{\boldsymbol{u}}, \boldsymbol{x}_{\boldsymbol{v}}) d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}} \tag{13}$$

Let $\delta(\boldsymbol{x}_{\boldsymbol{u}})$ be the Gâteaux's derivative at $\boldsymbol{x}_{\boldsymbol{u}}$, then

$$F(f_{\boldsymbol{u}} + t\delta_{\boldsymbol{u}}, g_{\boldsymbol{v}}) - F(f_{\boldsymbol{u}}, g_{\boldsymbol{v}}) = \int \left(2tf\delta + t^2\delta^2 - 2th\delta + 2tg\delta \right) p_{\boldsymbol{u}\boldsymbol{v}} d\boldsymbol{x}_{\boldsymbol{u}} d\boldsymbol{x}_{\boldsymbol{v}}$$

At minimum, we have

$$\begin{aligned}
&\lim_{t\to 0}\frac{F(f_{\boldsymbol{u}}+t\delta_{\boldsymbol{u}},g_{\boldsymbol{v}})-F(f_{\boldsymbol{u}},g_{\boldsymbol{v}})}{t}\\
&=\lim_{t\to 0}\int\left(2f\delta+t\delta^2-2h\delta+2g\delta\right)p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{v}}\\
&=\int\left(2f\delta-2h\delta+2g\delta\right)p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{v}}\\
&=0
\end{aligned}\tag{14}$$

Since (14) holds for all $\delta\in L^2_{\boldsymbol{u}}$, then we have

$$\int\left(f+g-h\right)p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}=0\tag{15}$$

By symmetry, we also have the following identity.

$$\int\left(f+g-h\right)p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}=0\tag{16}$$

Since h is given, we set $C_1=\int hp_{\boldsymbol{uv}}dv$ and $C_2=\int hp_{\boldsymbol{uv}}du$.

Solving (15) for f we have,

$$f=\frac{\int hp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}-\int gp_{\boldsymbol{uv}}d\boldsymbol{x}v}{\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}\tag{17}$$

Plug (17) into (16), we have

$$\begin{aligned}
&\int\frac{\int hp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}-\int gp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}{\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}+g\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}-\int hp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}=0\\
\Leftrightarrow\quad & g\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}-\int\frac{\int gp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}{\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}=\int hp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}-\int\frac{\int hp_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}{\int p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{v}}}p_{\boldsymbol{uv}}d\boldsymbol{x}_{\boldsymbol{u}}
\end{aligned}\tag{18}$$

Since $\boldsymbol{X}_{\boldsymbol{u}}\perp\boldsymbol{X}_{\boldsymbol{v}}$, we have $p_{\boldsymbol{uv}}=p_{\boldsymbol{u}}p_{\boldsymbol{v}}$. Then, identity (18) is equivalent to

$$g-\int gp_{\boldsymbol{v}}d\boldsymbol{x}_{\boldsymbol{v}}=\int hp_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{u}}-\int\int hp_{\boldsymbol{v}}d\boldsymbol{x}_{\boldsymbol{v}}p_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{u}}.$$

This is a Fredholm integral equation, with solution

$$\begin{cases}f=\int hp_{\boldsymbol{v}}d\boldsymbol{x}_{\boldsymbol{v}}-C\\ g=\int hp_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{u}}+C\end{cases}\tag{19}$$

where C is any constant.

To this end, the minimum approximation error in (12) achieves 0 when the following identity is true almost surely.

$$h=\int hp_{\boldsymbol{v}}d\boldsymbol{x}_{\boldsymbol{v}}+\int hp_{\boldsymbol{u}}d\boldsymbol{x}_{\boldsymbol{u}}$$

A counter example is given by $h(\boldsymbol{x}_{\boldsymbol{u}},\boldsymbol{x}_{\boldsymbol{v}})=\sin(\boldsymbol{x}_{\boldsymbol{u}}+\boldsymbol{x}_{\boldsymbol{v}})$ which does not assume the above decomposition. So $L^2_{\boldsymbol{u}}\oplus L^2_{\boldsymbol{v}}$ is a proper subspace of $h(\boldsymbol{x}_{\boldsymbol{u}},\boldsymbol{x}_{\boldsymbol{v}})\in L^2([0,1]^{|\boldsymbol{u}\cup\boldsymbol{v}|})$.

Thus the proposition is proved. □

Theorem 1. *Let $\mathcal{G}^a$ be the set of amiable group additive structures for f. There is a unique minimal group additive structure $G^*\in\mathcal{G}^a$ such that $G^*\leq G$ for all $G\in\mathcal{G}^a$, where the order is given by Definition 1. G^* is called the* ***intrinsic group additive structure*** *for f.*

Proof. Since the partial order is defined for any subset of group structures in $\mathcal{G}^a$, the existence of G^* is the result of Zorn's Lemma. The uniqueness is due to the fact that $\mathcal{G}^a$ is a finite set. □

Proposition 3. *Let G be a group additive structure and $\mathcal{H}_G$ be the induced direct sum RKHS defined in (3). Then, we have the following inequality relating the covering number of $\mathcal{H}_G$ and the covering numbers of $\mathcal{H}_{\boldsymbol{u}_j}$*

$$\ln \mathcal{N}(\epsilon, I_G, d_\infty) \leq \sum_{j=1}^{d} \ln \mathcal{N}\left(\frac{\epsilon}{|G|}, I_{\boldsymbol{u}_j}, d_\infty\right), \tag{20}$$

where $|G|$ denotes the number of groups in G.

Proof. Due to Lemma 1, we have $\mathcal{N}(\epsilon, I_G, d_\infty) \leq \Pi_{\boldsymbol{u}\in G}\mathcal{N}\left(\frac{\epsilon}{|G|}, \overline{I(\widetilde{\mathcal{H}}_{\boldsymbol{u}})}, d_\infty\right) = \Pi_{\boldsymbol{u}\in G}\mathcal{N}\left(\frac{\epsilon}{|G|}, I_{\boldsymbol{u}}, d_\infty\right)$. Then, taking log on both sides gives the desired result. □

Proposition 4. *Let G^* be the intrinsic group additive structure, $G \in \mathcal{G}^a$ a given amiable group structure, and $\mathcal{H}_{G^*}$ and $\mathcal{H}_G$ the respective direct sum RKHSs. If $\hat{f}_G^\lambda \in \mathcal{H}_G$ is the optimal solution of Problem (7), then for any $\epsilon > 0$, we have*

$$P\left(|\widehat{\mathcal{R}}(\hat{f}_G) - \mathcal{R}(f_{G^*}^*)| > \epsilon\right) \leq 12n \cdot \exp\left\{\sum_{\boldsymbol{u}\in G} \ln \mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, d_\infty\right) - \frac{\epsilon^2 n}{144}\right\} + 12n \cdot \exp\left\{\sum_{\boldsymbol{u}\in G} \ln \mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, d_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f_{G^*}^*\|^2}{12}\right)^2\right\}. \tag{21}$$

Proof. Since the following inequality holds,

$$|\widehat{\mathcal{R}}_g(\hat{f}_G) - \mathcal{R}_g(f_{G^*}^*)| \leq |\widehat{\mathcal{R}}_g(\hat{f}_G) - \mathcal{R}_g(\hat{f}_G)| + |\mathcal{R}_g(\hat{f}_G) - \mathcal{R}_g(f_{G^*}^*)|, \tag{22}$$

the upper bound for the desired deviation can be derived from the upper bounds of the two terms on RHS in the inequality.

The upper bound for the first term can be derived by using the uniform convergence bound in [1] (also see Lemma 12.38 in [20]). So we have the following probabilistic upper bound for the first term. For all $n > \frac{8}{\epsilon^2}$,

$$\begin{aligned} & P\left(|\widehat{\mathcal{R}}_g(\hat{f}_G) - \mathcal{R}_g(\hat{f}_G)| > \frac{\epsilon}{2}\right) \\ \leq & 12n \cdot \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \cdot \exp\left\{-\frac{\epsilon^2 n}{144}\right\} \\ \leq & 12n \cdot \exp\left\{\ln \mathcal{N}^{(n)}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - \frac{\epsilon^2 n}{144}\right\} \\ \leq & 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - \frac{\epsilon^2 n}{144}\right\}, \end{aligned} \tag{23}$$

where $\ell_\infty^{X'}$ denotes the sup-norm of function $f \in \mathcal{F}$ restricted to the sample $X' = \{x_1', \ldots, x_n'\}$ which is independent of the sample $X = \{x_1, \ldots, x_n\}$ used for estimation and $\mathcal{N}^{(n)}(\epsilon, \mathcal{H}, \ell_\infty)$ is called the *ϵ-growth function* of the space $\mathcal{H}$ which is defined as

$$\mathcal{N}^{(n)}(\epsilon, \mathcal{H}, \ell_\infty) := \sup_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}} \mathcal{N}\left(\epsilon, \mathcal{H}, \ell_\infty^X\right).$$

The second inequality is due to the fact that $\mathbb{E}\left[\mathcal{N}\left(\epsilon, \mathcal{H}, \ell_\infty^{X'}\right)\right] \leq \mathcal{N}^{(n)}(\epsilon, \mathcal{H})$.

The upper bound for the second term in 22 can be derived by repeatedly applying the same uniform convergence bound. Due to Lemma 2, we have for all $\epsilon > 0$ and all $n > 2/\epsilon^2$,

$$\begin{aligned}
&P\left(|\mathcal{R}_g(\hat{f}_G) - \mathcal{R}_g(f^*_{G^*})| > \frac{\epsilon}{2}\right) \\
\leq& 12n \cdot \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \cdot \exp\left\{-n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\} \\
\leq& 12n \cdot \exp\left\{\ln \mathcal{N}^{(n)}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\} \\
\leq& 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\}. \qquad (24)
\end{aligned}$$

By plugging the upper bounds (23) and (24) in (22), we have

$$\begin{aligned}
&P\left(|\widehat{\mathcal{R}}_g(\hat{f}_G) - \mathcal{R}_g(f^*_{G^*})| > \epsilon\right) \\
\leq& 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - \frac{\epsilon^2 n}{144}\right\} + \\
&\qquad 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\} \qquad (25)
\end{aligned}$$

By using Lemma 3, we can bound the covering number for $\mathcal{H}_G$ from above and obtain the following inequality.

$$\begin{aligned}
P\left(|\widehat{\mathcal{R}}_g(\hat{f}_G) - \mathcal{R}_g(f^*_{G^*})| > \epsilon\right) \leq& 12n \cdot \exp\left\{\sum_{\boldsymbol{u} \in G} \ln \mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, \ell_\infty\right) - \frac{\epsilon^2 n}{144}\right\} + \\
&12n \cdot \exp\left\{\sum_{\boldsymbol{u} \in G} \ln \mathcal{N}\left(\frac{\epsilon}{12|G|}, \mathcal{H}_{\boldsymbol{u}}, \ell_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\}.
\end{aligned}$$

□

Theorem 2. *Let $\mathcal{G}^a$ be the set of all amiable group structures. For any $\epsilon > 0$ and $n > 2/\epsilon^2$, we have*

$$\begin{aligned}
P\left(\sup_{G \in \mathcal{G}^a} |\widehat{\mathcal{R}}_g(\hat{f}^\lambda_G) - \mathcal{R}_g(f^*_{G^*})| > \epsilon\right) \leq& 12n|\mathcal{G}^a| \cdot \left[\exp\left\{\max_{G \in \mathcal{G}^a} \ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, d_\infty\right) - \frac{\epsilon^2 n}{144}\right\}\right. \\
&\left. + \exp\left\{\max_{G \in \mathcal{G}^a} \ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, d_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f^*_{G^*}\|^2}{12}\right)^2\right\}\right] \qquad (26)
\end{aligned}$$

Proof. Denote $\mathcal{D}^{(n)}_{G,\epsilon} = \left\{(\boldsymbol{x}_i, y_i)_{i=1}^n \in \mathcal{X} \times \mathcal{Y} \middle| |\widehat{\mathcal{R}}(\hat{f}_G, g) - \mathcal{R}(f^*_{G^*}, g)| > \epsilon\right\}$, then we have

$$\begin{aligned}
P\left(\bigcup_{G \in \mathcal{G}^a} \mathcal{D}_{G,\epsilon}\right) \leq& \sum_{G \in \mathcal{G}^a} P\left(\mathcal{D}_{G,\epsilon}\right) \\
&\leq |\mathcal{G}^a| 12n \exp\left\{\max_{G \in \mathcal{G}^a} \ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - \frac{\epsilon^2 n}{144}\right\} + \\
&\qquad |\mathcal{G}^a| 12n \exp\left\{\max_{G \in \mathcal{G}^a} \ln \mathcal{N}\left(\frac{\epsilon}{12}, \mathcal{H}_G, \ell_\infty\right) - n\left(\frac{\epsilon}{24} - \frac{\lambda \|f^*_{G^*}\|^2}{12}\right)^2\right\}
\end{aligned}$$

where the second inequality is due to the proof of Proposition 4. □

Theorem 3. *(i) For a non-amiable group structure $G' \in \mathcal{G} \setminus \mathcal{G}^a$, there exists a constant $C > 0$ such that $|\widehat{\mathcal{R}}_g(\hat{f}^\lambda_{G'}) - \mathcal{R}_g(f^*_{G^*})|$ converges to C in probability. (ii) There exits a constant $\tilde{C}$ such that $P(|\widehat{\mathcal{R}}_g(\hat{f}^\lambda_{G'}) - \mathcal{R}_g(f^*_{G^*})| > \tilde{C}$ for all $G' \in \mathcal{G} \setminus \mathcal{G}^a)$ goes to 1 as n goes to infinity.*

Proof. We start with the following triangle inequality

$$|\widehat{\mathcal{R}}_g(\hat{f}_{G'}) - \mathcal{R}_g(f^*_{G^*})| \leq |\widehat{\mathcal{R}}_g(\hat{f}_{G'}) - \mathcal{R}_g(\hat{f}_{G'})| + |\mathcal{R}_g(\hat{f}_{G'}) - \mathcal{R}_g(f^*_{G^*})|. \tag{27}$$

The first term on the RHS can be bounded by using the same uniform convergence bound (12.135) in [20]. For any $\epsilon > 0$ and all $n > 2/\epsilon^2$,

$$\begin{aligned}\mathbb{P}\left(|\widehat{\mathcal{R}}_g(\hat{f}_{G'}) - \mathcal{R}_g(\hat{f}_{G'})| > \epsilon\right) &\leq 12n \cdot \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon}{6}, \mathcal{H}_{G'}, \ell_\infty^{X'}\right)\right] \exp\left\{-\frac{\epsilon^2 n}{36}\right\} \\ &\leq 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{6}, \mathcal{H}_{G'}, \ell_\infty\right) - \frac{\epsilon^2 n}{36}\right\}. \end{aligned} \tag{28}$$

In order to derive an upper bound for the second term, we first decompose each risk into bias and variance. According to [6], the risk of the empirical estimate of $\hat{f}_{G'}$ can be decomposed as

$$\begin{aligned}\mathcal{R}_g(f_{G'}) &= \int_{\mathcal{X}\times\mathcal{Y}} \left(g(y) - \hat{f}_{G'}(x)\right)^2 dP_{XY} \\ &= \int_{\mathcal{X}\times\mathcal{Y}} \left(g(y) - f_{Y|X}(x)\right)^2 dP_{XY} + \int_{\mathcal{X}\times\mathcal{Y}} \left(f_{X|Y}(x) - \hat{f}_{G'}(x)\right)^2 dP_{XY}, \end{aligned} \tag{29}$$

where $f_{X|Y}(x) := \int_{\mathcal{Y}} g(y) dP_{Y|X}$ is the optimal regression function.

By assuming $f_{X|Y}(x) = f^*_{G^*}$ (this is the assumption we use throughout this chapter), we have

$$|\mathcal{R}_g(\hat{f}_{G'}) - \mathcal{R}_g(f^*_{G^*})| = \int_{\mathcal{X}\times\mathcal{Y}} \left(f^*_{G^*}(x) - \hat{f}_{G'}(x)\right)^2 dP_{XY} \tag{30}$$

According to Theorem 2.1 in [13], we have the following decompositions for the two function on the RHS of (30):

$$\begin{aligned} f^*_{G^*} &= \sum_{\boldsymbol{u}\subseteq\{1,\ldots,p\}} f^*_{G^*,\boldsymbol{u}} \quad \text{with} \quad f^*_{G^*,\boldsymbol{u}} := \sum_{\boldsymbol{v}\subseteq\boldsymbol{u}} (-1)^{|\boldsymbol{u}|-|\boldsymbol{v}|}\, \mathrm{P}_{\{1,\ldots,p\}\setminus\boldsymbol{v}}(f^*_{G^*}), \\ \hat{f}_{G'} &= \sum_{\boldsymbol{u}\subseteq\{1,\ldots,p\}} \widehat{f}_{G',\boldsymbol{u}} \quad \text{with} \quad \hat{f}_{G',\boldsymbol{u}} := \sum_{\boldsymbol{v}\subseteq\boldsymbol{u}} (-1)^{|\boldsymbol{u}|-|\boldsymbol{v}|}\, \mathrm{P}_{\{1,\ldots,p\}\setminus\boldsymbol{v}}(\widehat{f}_{G'}). \end{aligned}$$

Since G' is an non-amiable group structure, there is at least one subset[1] of $\boldsymbol{u} \subseteq \{1,\ldots,p\}$ such that $f^*_{G^*,\boldsymbol{u}} \neq \hat{f}_{G',\boldsymbol{u}}$. Let $C = \min_{\boldsymbol{u}\subseteq\{1,\ldots,p\}} \int_{\mathcal{X}\times\mathcal{Y}} \left(f^*_{G^*,\boldsymbol{u}} - \hat{f}_{G',\boldsymbol{u}}\right)^2 dP_{XY} > 0$ and denote $\boldsymbol{u}^c = \{1,\ldots,p\} \setminus \boldsymbol{u}$, then we have

$$\begin{aligned} &\int_{\mathcal{X}\times\mathcal{Y}} \left(f^*_{G^*}(\boldsymbol{x}) - \hat{f}_{G'}(\boldsymbol{x})\right)^2 dP_{XY} \\ =&\int_{\mathcal{X}_{\boldsymbol{u}}\times\mathcal{Y}} \left(f^*_{G^*,\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}}) - \hat{f}_{G',\boldsymbol{u}}(\boldsymbol{x}_{\boldsymbol{u}})\right)^2 dP_{X_{\boldsymbol{u}}Y} + \int_{\mathcal{X}_{\boldsymbol{u}^c}\times\mathcal{Y}} \left(f^*_{G^*,\boldsymbol{u}^c}(\boldsymbol{x}_{\boldsymbol{u}^c}) - \hat{f}_{G',\boldsymbol{u}^c}(\boldsymbol{x}_{\boldsymbol{u}^c})\right)^2 dP_{X_{\boldsymbol{u}^c}Y} \\ \geq& C + \int_{\mathcal{X}\times\mathcal{Y}} \left(f^*_{G^*,\boldsymbol{u}^c}(\boldsymbol{x}_{\boldsymbol{u}^c}) - \hat{f}_{G',\boldsymbol{u}^c}(\boldsymbol{x}_{\boldsymbol{u}^c})\right)^2 dP_{XY} \\ >& 0. \end{aligned} \tag{31}$$

where the first equality is due to the orthogonality possessed by a direct sum Hilbert space.

By using (27), (28), (30) and (31), we can obtain

$$P\left(|\widehat{\mathcal{R}}_g(\hat{f}_{G'}) - \mathcal{R}_g(f^*_{G^*})| > \epsilon + C\right) \leq 12n \cdot \exp\left\{\ln \mathcal{N}\left(\frac{\epsilon}{6}, \mathcal{H}_{G'}, \ell_\infty\right) - \frac{\epsilon^2 n}{36}\right\} \tag{32}$$

□

[1] If G' is amiable, then a subset $\boldsymbol{u}$ of G' always assumes an additive structure. So there is no error between $f^*_{G^*}$ and $\hat{f}_{G'}$ after such a decomposition.

Theorem 4. *Let $\lambda_n * n \to 0$. By choosing a proper tuning parameter $\mu > 0$ for the structural penalty, the estimated group structure $\hat{G}$ is consistent for the intrinsic group additive structure G^*, that is, $P(\hat{G} = G^*)$ goes to one as the sample size n goes to infinity.*

Proof. According to Theorem 3, by choosing $\epsilon < C$, an agreeable group structure will be chosen with high probability.

For an amiable group structure, let $\epsilon_1 = |\hat{\mathcal{R}}_g(\hat{f}_G^{\lambda}) - \mathcal{R}_g(f_{G^*}^*)|$ and $\epsilon_2 = \mu\boldsymbol{\mathcal{C}}(G) - \mu\boldsymbol{\mathcal{C}}(G^*)$. Since $\boldsymbol{\mathcal{C}}(G) > \boldsymbol{\mathcal{C}}(G^*)$ when G is not the true group structure, we have $\epsilon_2 > 0$. Because ϵ_1 converges to 0 in probability. Thus the true group structure G^* will be picked with high probability if Problem (7) is solved. □

Lemma 1. *Let $S, T : \mathcal{F}_1 \to \mathcal{F}_2$ be operators in Banach spaces and $\epsilon_1, \epsilon_2 > 0$. Then we have*

$$\mathcal{N}(\epsilon_1 + \epsilon_2, T + S) \leq \mathcal{N}(\epsilon_1, S) \cdot \mathcal{N}(\epsilon_2, T).$$

Lemma 2. *For all $\epsilon > 0$ and all $n > 2/\epsilon^2$,*

$$P\left(|\mathcal{R}_g\left(\hat{f}_G\right) - \mathcal{R}_g\left(f_{G^*}^*\right)| > \frac{\epsilon}{2}\right) \leq 12n \cdot \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \exp\left\{-n\left(\frac{\epsilon}{24} - \frac{\lambda_n \|f_{G^*}^*\|^2}{12}\right)^2\right\}.$$

Proof. Due to the uniform convergence bound (12.135) in [20], given $\widehat{f}_G$,, we have for all $\epsilon > 0$ and all $n \geq 2/\epsilon^2$,

$$P\left(|\widehat{\mathcal{R}}_g\left(\hat{f}_G\right) - \mathcal{R}_g\left(\hat{f}_G\right)| > \epsilon\right) \leq 12n \cdot \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \exp\left\{-\frac{\epsilon^2 n}{36}\right\}.$$

By setting $\delta = 12n \cdot \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \exp\left\{-\frac{\epsilon^2 n}{36}\right\}$ and solve for ϵ, we have

$$\epsilon = 6n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2}.$$

Equivalently with probability at least $1 - \delta$,

$$|\widehat{\mathcal{R}}_g\left(\hat{f}_G\right) - \mathcal{R}_g\left(\hat{f}_G\right)| \leq 6n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2}.$$

Due to the symmetry of the above bound, we have with probability at least $1 - \delta$,

$$\begin{aligned}
\mathcal{R}_g\left(\hat{f}_G\right) &\leq \widehat{\mathcal{R}}_g\left(\hat{f}_G\right) + 6n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2} \\
&\leq \widehat{\mathcal{R}}_g\left(\hat{f}_G\right) + \lambda\|\hat{f}_G\|^2 + 6n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2} \\
&\leq \widehat{\mathcal{R}}_g\left(f_{G^*}^*\right) + \lambda\|f_{G^*}^*\|^2 + 6n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2} \\
&\leq \mathcal{R}_g\left(f_{G^*}^*\right) + \lambda\|f_{G^*}^*\|^2 + 12n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2}
\end{aligned}$$

where the third inequality is due to the definition of $\hat{f}_G$ as the minimizer of the empirical problem. We applied the uniform convergence bound twice, one for the first inequality and the other for the last inequality.

Since it is always true that $\mathcal{R}_g\left(f_{G^*}^*\right) \leq \mathcal{R}_g\left(\hat{f}_G\right)$, we have the symmetric upper bound with probability $1 - \delta$,

$$\begin{aligned}
&|\mathcal{R}_g\left(\hat{f}_G\right) - \mathcal{R}_g\left(f_{G^*}^*\right)| \leq \\
&\qquad \lambda\|f_{G^*}^*\|^2 + 12n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln \delta\right)^{1/2}.
\end{aligned}$$

By setting $\lambda\|f^*_{G^*}\|^2 + 12n^{-1/2}\left(\ln 12n + \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] - \ln\delta\right)^{1/2} = \epsilon/2$ and solve for δ, we have

$$\delta = 12n \cdot \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \exp\left\{-n\left(\frac{\epsilon}{24} - \frac{\lambda_n\|f^*_{G^*}\|^2}{12}\right)^2\right\}$$

Thus the bound for the second term is for all $\epsilon > 0$ and all $n > 2/\epsilon^2$,

$$P\left(|\mathcal{R}_g\left(\hat{f}_G\right) - \mathcal{R}_g\left(f^*_{G^*}\right)| > \frac{\epsilon}{2}\right) \leq$$
$$12n \cdot \ln \mathbb{E}\left[\mathcal{N}\left(\frac{\epsilon^2}{12}, \mathcal{H}_G, \ell_\infty^{X'}\right)\right] \exp\left\{-n\left(\frac{\epsilon}{24} - \frac{\lambda_n\|f^*_{G^*}\|^2}{12}\right)^2\right\}$$

□

The following Lemma is taken from Lemma 1 in [12], which shows the relationship between the covering number of the direct sum of two operators and the covering numbers of the individual operators.

Lemma 3. *Let $S, T : \mathcal{B}_1 \to \mathcal{B}_2$ be operators in real Banach spaces and $\epsilon, \delta > 0$. Then,*

$$\mathcal{N}\left(\epsilon + \delta, T + S\right) \leq \mathcal{N}\left(\epsilon, T\right) \cdot \mathcal{N}\left(\delta, S\right).$$

B. Complete Simulation Results

μ	α	M1	M2	M3	M4	M5
1.0000e-10	1.00	0	0	0	0	100
1.0000e-10	2.00	0	0	0	0	100
1.0000e-10	3.00	0	0	0	0	100
1.0000e-10	4.00	0	0	0	0	99
1.0000e-10	5.00	0	0	0	0	10
1.0000e-10	6.00	0	0	0	0	0
1.0000e-10	7.00	0	0	0	0	0
1.0000e-10	8.00	0	0	0	0	0
1.0000e-10	9.00	0	0	0	0	0
1.0000e-10	10.00	0	0	0	0	0
1.1180e-08	1.00	0	0	0	0	100
1.1180e-08	2.00	0	0	0	0	98
1.1180e-08	3.00	0	0	0	0	0
1.1180e-08	4.00	0	0	0	0	0
1.1180e-08	5.00	0	0	0	0	0
1.1180e-08	6.00	0	0	0	0	0
1.1180e-08	7.00	0	0	0	0	0
1.1180e-08	8.00	0	0	0	1	0
1.1180e-08	9.00	0	0	0	77	0
1.1180e-08	10.00	0	0	0	92	0
1.2500e-06	1.00	0	0	0	0	100
1.2500e-06	2.00	0	0	0	0	0
1.2500e-06	3.00	14	0	0	84	0
1.2500e-06	4.00	81	3	4	99	0
1.2500e-06	5.00	90	77	77	99	0
1.2500e-06	6.00	94	92	90	99	0
1.2500e-06	7.00	96	96	95	100	0
1.2500e-06	8.00	98	97	96	100	0
1.2500e-06	9.00	98	97	97	100	0
1.2500e-06	10.00	100	97	97	100	0
1.3975e-04	1.00	0	0	0	0	100
1.3975e-04	2.00	0	95	93	100	0
1.3975e-04	3.00	100	95	92	90	0
1.3975e-04	4.00	100	28	23	9	0
1.3975e-04	5.00	100	13	12	3	0
1.3975e-04	6.00	100	5	7	3	0
1.3975e-04	7.00	100	0	0	2	0
1.3975e-04	8.00	100	0	0	0	0
1.3975e-04	9.00	100	0	0	0	0
1.3975e-04	10.00	100	0	0	0	0
1.5625e-02	1.00	0	0	0	0	100
1.5625e-02	2.00	0	0	0	100	0
1.5625e-02	3.00	100	0	0	0	0
1.5625e-02	4.00	100	0	0	0	0
1.5625e-02	5.00	100	0	0	0	0
1.5625e-02	6.00	100	0	0	0	0
1.5625e-02	7.00	100	0	0	0	0
1.5625e-02	8.00	100	0	0	0	0
1.5625e-02	9.00	100	0	0	0	0
1.5625e-02	10.00	100	0	0	0	0

Table 3: Frequencies that the true group structures are selected under different parameter pairs for the five models. Exhaustive search algorithm without parameter turning.

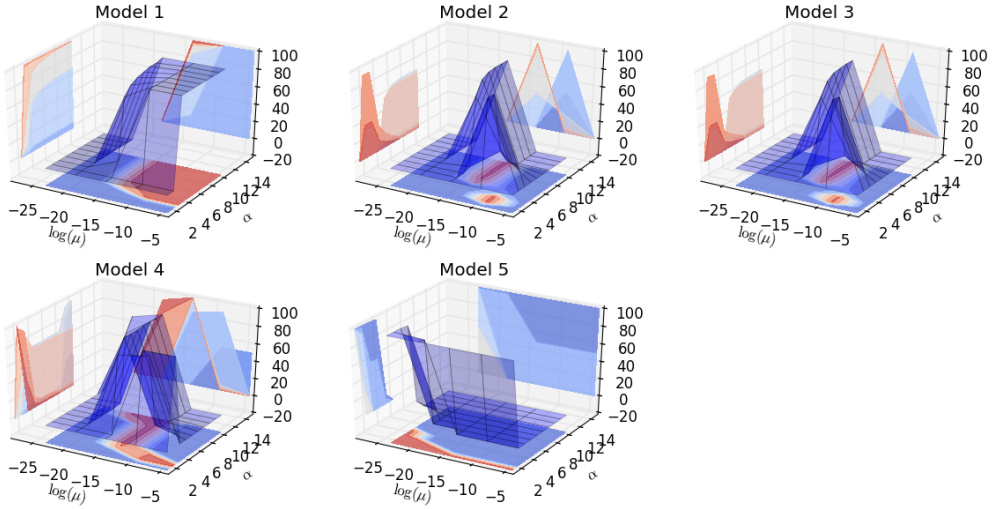

Figure 3: The 3D surface of the frequencies (out of 100) that the true group structures are identified for the five chosen models in Table 1 over the entire parameters grid. Given a (μ, α) pair, the penalized goodness of fit is calculated for all group structures. We recorded each time the true group structure is identified. The values of μ are reported in log-scale. Each surface plot is accompanied with three contour plots as the 2D projections of the surface to enhance the effect of the visualization.

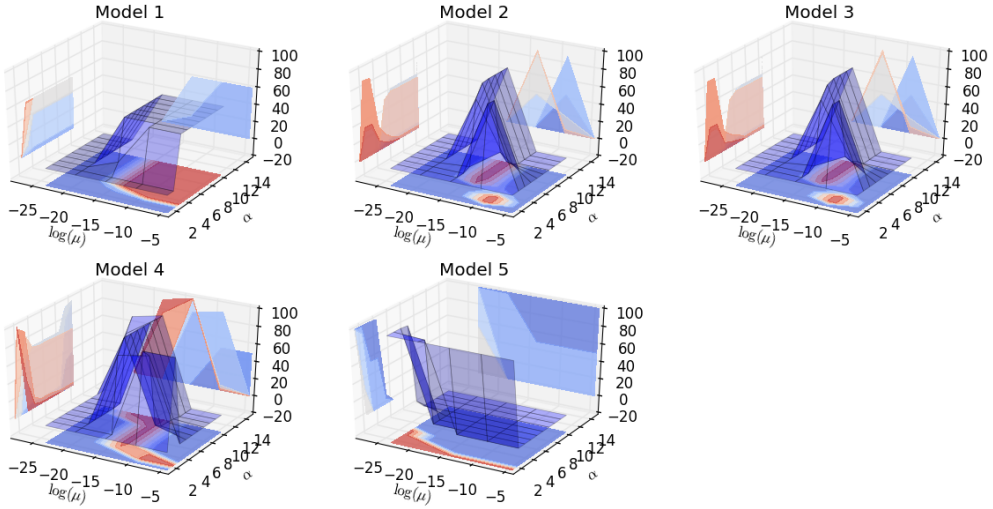

Figure 4: The 3D surfaces of the frequencies (out of 100) that the true group structures are identified for the five chosen models in Table 1 over the entire parameter grids. The training procedure uses a separate validation data set to select the optimal tuning parameters (μ, α). The values of μ are reported in log-scale. Each surface plot is accompanied with three contour plots as the 2D projections of the surface to enhance the effect of the visualization.

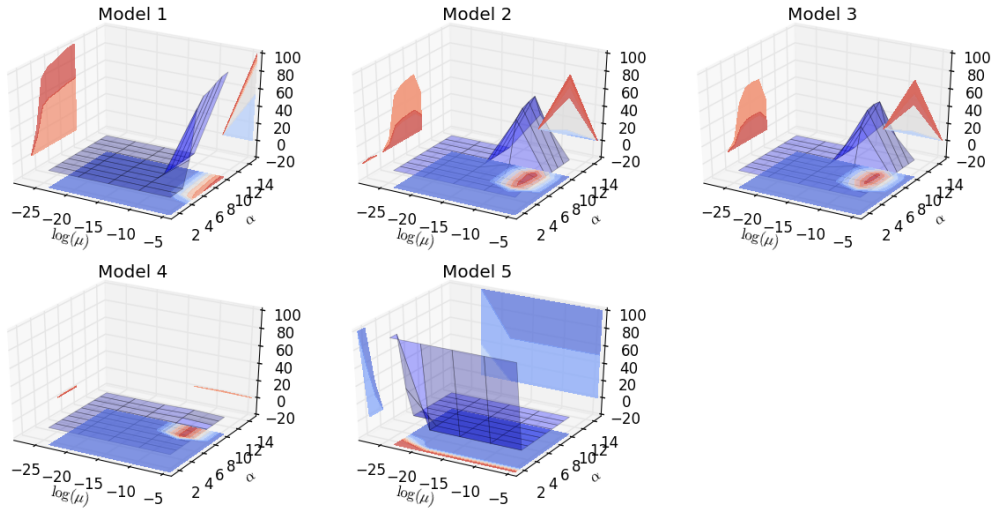

Figure 5: The 3D surfaces of the frequencies (out of 100) that the true group structures are identified for the five chosen models in Table 1 over the entire parameter grids. The training uses the backward stepwise algorithm and the procedure uses a separate validation data set to select the optimal tuning parameters (μ, α). The values of μ are reported in log-scale. Each surface plot is accompanied with three contour plots as the 2D projections of the surface to enhance the effect of the visualization.

C. Complete Results of Real Data Applications

Application 1: Boston Housing Data

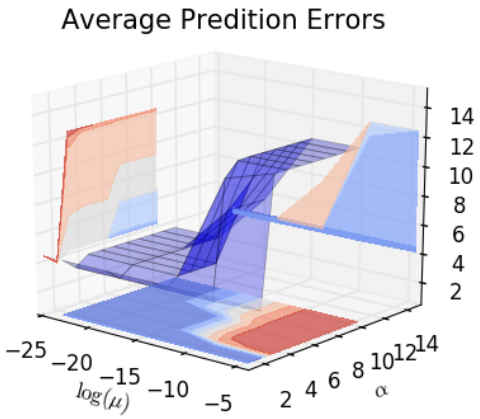

Figure 6: The results of applying the backward step-wise algorithm on Boston Housing data with 10-fold CV. The 3D surfaces shows the average validation error over the entire grid of (μ, α) pairs. The surface plot is accompanied with three contour plots as the 2D projections of the surface to enhance the effect of the visualization.

Application 2: Communities and Crime Data

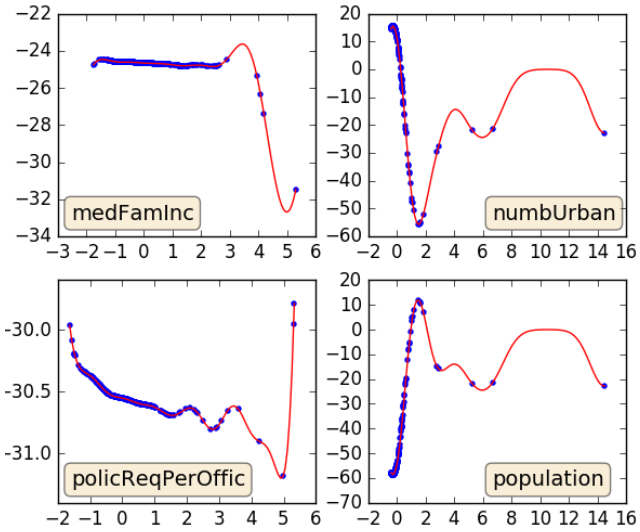

Figure 7: Selected results for the communities and crime data where the number of murders is the response. The blue dots are the transformed observation of the predictor variable. The red line is the estimated function.