Taylor & Francis
Taylor & Francis Group

Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

Nonparametric Prediction Distribution from Resolution-Wise Regression with Heterogeneous Data

Jialu Li, Wan Zhang, Peiyao Wang, Qizhai Li, Kai Zhang & Yufeng Liu

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 00, NO. 0, 1–16
https://doi.org/10.1080/07350015.2022.2115498

Taylor & Francis
Taylor & Francis Group

Nonparametric Prediction Distribution from Resolution-Wise Regression with Heterogeneous Data

Jialu Li*[a], Wan Zhang*[b], Peiyao Wang[b], Qizhai Li[c], Kai Zhang[b], and Yufeng Liu[d]

[a]School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, China; [b]Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, NC.; [c]LSC, NCMIS, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, and University of Chinese Academy of Sciences, Beijing, China; [d]Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Science, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC

ABSTRACT

Modeling and inference for heterogeneous data have gained great interest recently due to rapid developments in personalized marketing. Most existing regression approaches are based on the conditional mean and may require additional cluster information to accommodate data heterogeneity. In this article, we propose a novel nonparametric resolution-wise regression procedure to provide an estimated distribution of the response instead of one single value. We achieve this by decomposing the information of the response and the predictors into resolutions and patterns, respectively, based on marginal binary expansions. The relationships between resolutions and patterns are modeled by penalized logistic regressions. Combining the resolution-wise prediction, we deliver a histogram of the conditional response to approximate the distribution. Moreover, we show a sure independence screening property and the consistency of the proposed method for growing dimensions. Simulations and a real estate valuation dataset further illustrate the effectiveness of the proposed method.

ARTICLE HISTORY
Received October 2020
Accepted July 2022

KEYWORDS
Binary expansion; Data heterogeneity; Nonparametric statistics; SSANOVA; Sure independence screening

1. Introduction

A common nonparametric regression model establishes the effects of the explanatory variables on the response variable in the form of

$$Y = f(X) + \varepsilon, \qquad (1)$$

where Y is the response variable, $X = (X_1, \ldots, X_q)^T$ is the q-dimensional explanatory variable vector, and ε is the random error, which is often assumed with mean 0 and variance σ^2, and is independent of X.

In recent years there has been a growing demand for exploring regression methods for heterogeneous populations, which has broad applications in personalized marketing and other fields. One characteristic of data heterogeneity is the existence of subpopulations in the data. In practice, the heterogeneity can be regarded as the result of some latent variables. This happens frequently since it is difficult to collect all the explanatory variables for the response. For example, in the real estate data in Section 6, a river and a highway through the city create subpopulations and heterogeneous distributions of housing prices. However, the information of this river and this highway is not available in the data.

Denote the unobserved categorical variable by Z taking values $1, \ldots T$, where T is unknown. Suppose the potential true relationship of the response and all the explanatory variables can be expressed by

$$Y = \sum_{t=1}^{T} f_t(X) I(Z = t) + \varepsilon, \qquad (2)$$

with unknown functions f_t's, $t = 1, \ldots, T$. In this article, our goal is to relate Y with X without knowing Z. However, this differs from fitting model (1), since the true relationship between Y and X may not even be a function. As an illustration, the housing prices on the two sides of Tamsui river with respect to longitude and latitude are shown in Figure 1. The plot shows a mixture of two subgroups: the housing prices on the west and the east of the river behave differently. The latent variable Z, that is, the indicator for which side the position on, determines the two subgroups. Without knowing Z, the relationship between Y and X cannot be captured by a single function. Hence, new methods to model the effects of X on Y with such a challenging heterogeneous population are in great needs.

One possible idea is to use smoothing splines (Green and Silverman 1994) which can capture local behaviors. A more general setting is the smoothing spline analysis of variance (SSANOVA) (Wahba 1990; Gu 2002), which fits an additive model for main effects and interactions. These approaches use regression to estimate the overall conditional mean function,

CONTACT Yufeng Liu yfliu@email.unc.edu Department of Statistics and Operations Research, Department of Genetics, Department of Biostatistics, Carolina Center for Genome Science, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA.

*These authors are co-first authors.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

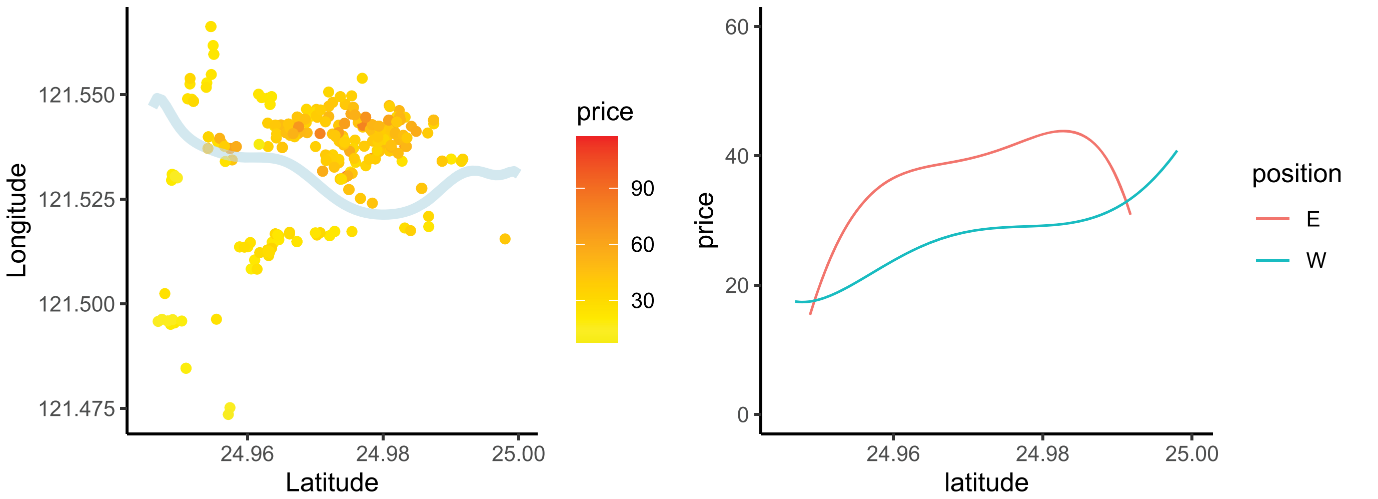

Figure 1. Left panel: An illustration of heterogeneous data in the housing prices on two sides of Tamsui River (light blue curve). Right panel: The housing prices follow different distributions: The housing prices on the west monotonically increase with latitude, while those on the east are concave and parabolic.

which tends to fit a compromised effect of the subgroups. Hence, they may fail to identify the subgroups of the population, and the results might not be informative for either of the subgroups. Moreover, the estimated distribution might not really reflect the pattern of the true one.

Another possible strategy is to cluster the data first, then fit a regression model within each subgroup. Existing model-based clustering approaches include Jacobs et al. (1991), Pan and Shen (2006), Raftery and Dean (2006), and Guo et al. (2010). As alternative approaches, Lindsten, Ohlsson, and Ljung (2011), Hocking et al. (2011), and Pan, Shen, and Liu (2013) formulated clustering as a penalized regression problem with fusion-type penalties. However, these methods focus on finding the groups based on the similarity of the explanatory variables, instead of identifying groups with different effects on the response.

In the literature, some individualized methods were proposed to handle heterogeneity. Ma and Huang (2017) employed subject-specific intercepts to model the unobserved factors which leads to the heterogeneity. They used a concave pairwise fusion penalty to shrink some intercepts to be the same, which can produce a partition of subgroups. Chen et al. (2021) considered a more general fuzing method than that of Ma and Huang (2017) to identify subgroups. Tang and Qu (2017) proposed a multi-directional penalty to shrink individuals to different groups. The performance of these methods depends on how well the subgroups are separated. If the subgroups are close to each other, the performance can be less accurate.

In this article, we tackle the heterogeneity from a new perspective. Instead of estimating f_t's in (2) through nonparametric regressions, we propose to estimate the conditional distribution of the response variable given observed explanatory variables. The estimated distribution provides an overall picture of the response variable, and can indicate the heterogeneity by the modes of the probability density function (PDF). To achieve this goal, one single regression is not enough, because the pattern of two or more possible values of the response corresponding to one observation of predictor variables cannot be expressed by an explicit function. Our idea is to consider binary expansion statistics proposed in Zhang (2019) and to decompose the response variable into several resolutions which can capture the local information. By establishing a set of logistic regressions, we relate the resolution information to the predictors. The set of regressions can model the heterogeneity since various estimations can be obtained from different local logistic regression models. To achieve the localization, we decompose the response variable by marginal binary expansions, which provides a balanced design and orthogonal resolutions. Our method eventually estimates the distribution of the response variable by a histogram, which shows the possible heterogeneity and even more complicated distributions, without any assumption of subgroup patterns. We show that the method has a sure independence screening property (Fan and Lv 2008; Fan and Song 2010) and provides consistent estimates for cell probabilities of the histogram for growing dimensions.

The rest of this article is organized as follows. In Section 2, we introduce resolution-wise regression, including the decomposition of the response variable and the establishment of the logistic regressions. Section 3 extends the proposed method to high-dimensional settings. In Section 4, we show the consistency of the estimated histogram. In Sections 5 and 6, we demonstrate the performance of our method by the simulated data and the real estate valuation dataset. Section 7 concludes this article. Some technical proofs and additional simulation results are presented in the Appendix and supplementary materials.

2. Methodology

A distribution estimation provides more information than a point estimation for heterogeneous data, as the estimated distribution can identify the subgroups by the shape of the PDF. For the case that the subpopulations are not obviously distinguishable from each other, the estimated distribution can still reflect the dispersion of the data.

A direct idea is splitting the range of Y by a partition $\min(Y) = a_0 \leqslant a_1 \leqslant \cdots \leqslant a_B = \max(Y)$ and modeling the probability of Y falling into each interval with X. This idea handles the heterogeneity by decomposing the information of Y into several nonoverlapping intervals. These intervals capture the local information of Y and work together to show the whole histogram. However, a drawback of this approach is the possible loss of information from negligence of joint information in two intervals and insufficient samples in each interval. In this article, we propose to construct overlapped resolutions based on binary expansions, where each resolution groups the distribution information of the union of several intervals, and includes all corresponding samples. In essence, the proposed construction leads to a balanced design and has a nonredundant orthogonality property. A histogram can be obtained by a transformation from resolution probabilities to cell probabilities. Here we refer to a cell as a bin of the histogram. In this section, we consider the one-dimensional X, and then extend to the high-dimensional case in Section 3. The rest of this section is organized as follows. In Section 2.1, we introduce the construction of the resolutions. In Section 2.2, a set of resolution-wise penalized logistic regressions are established. In Section 2.3, we introduce the binary interaction design (BID) equation to accomplish the transformation from the frequency domain to the probability domain.

2.1. Frequency Domain from Binary Expansions

To overcome the imbalance of nonoverlapping intervals, we use a balanced design based on binary expansions. A classical result on the binary expansion of a uniform random variable (Kac 1959) is given as follows:

Lemma 1. For $U \sim \text{Uniform}[0,1]$, we have $U = \sum_{k=1}^{\infty} \frac{V_k}{2^k}$, where $V_1, V_2, \dots, V_k, \dots$ iid follow Bernoulli(1/2).

Denote the cumulative distribution function (CDF) transformation of Y by U_Y. By Lemma 1, we have

$$U_Y = \sum_{k=1}^{\infty} \frac{B_k}{2^k},\ B_1, B_2, \dots, B_k, \dots \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2). \qquad (3)$$

Through the expansion, the information of Y is decomposed into the information of B_k's. In (3), binary variables B_k's can be regarded as indicator functions: $B_k = I(U_Y \in [\frac{1}{2^k}, \frac{2}{2^k}) \cup [\frac{3}{2^k}, \frac{4}{2^k}) \cup \dots \cup [\frac{2^k-1}{2^k}, 1]);\ k = 1, 2, \dots$

Figure 2 shows the binary variables B_1, B_2, B_3, respectively, with respect to U_Y. As a finite approximation of the infinite binary expansion, we can truncate the binary expansion of U_Y up to the d_Yth order

$$U_Y = \sum_{k=1}^{d_Y} \frac{B_k}{2^k}. \qquad (4)$$

Now we introduce the notations of resolutions. Using the binary variables taking values $\{-1, 1\}$ instead of $\{0, 1\}$ by the transformation $\dot{B}_k = 2B_k - 1$, the interactions of B_k's can be written as products. For example, the event $\{B_1 = 1, B_2 = 1\} \cup \{B_1 = 0, B_2 = 0\}$ is equivalent to $\{\dot{B}_1 \dot{B}_2 = 1\}$. In the remainder of this article, we shall use $\dot{B}_k \in \{-1, 1\}$.

To approximate the information given by Y, say the σ-field $\sigma(Y)$, we can use the σ-field generated by $\dot{B}_k$'s, denoted by $\sigma(\dot{B}_1, \dots, \dot{B}_{d_Y})$. For the truncation up to the d_Yth order, we can find a basis with $2^{d_Y} - 1$ variables

$$W_{\dot{B}} = \{\dot{B}_1, \dots, \dot{B}_{d_Y}, \dot{B}_1\dot{B}_2, \dots, \dot{B}_{d_Y-1}\dot{B}_{d_Y}, \dots, \prod_{i=1}^{d_Y} \dot{B}_i\}. \qquad (5)$$

We shall refer to the binary variables in $W_{\dot{B}}$ as resolutions of Y, and the set of all possible values of these resolutions as the frequency domain. Figure 3 shows the variables in $\sigma(\dot{B}_1, \dot{B}_2)$ with U_Y expanded up to the second order. Through this resolution decomposition, each variable takes value one on half of $[0, 1]$, and value negative one on the other half.

2.2. Logistic Regression in the Frequency Domain

With the resolutions decomposed from the binary expansion, we aim to model the relationship between each resolution and the predictors. The resolutions constructed by the binary expansion are independent with each other, thus, they can be modeled marginally. Since the resolutions are binary, it is essentially a classification problem. Note that for every resolution, Y is divided into two classes with groups of intervals according to the sign of the binary interaction. Hence, the decision boundary can be nonlinear. Therefore, we propose to use binary expansions of predictors as a nonparametric basis to fit a logistic regression on each resolution. Similar to the construction of U_Y, we have the binary expansion of U_X, up to the d_Xth order to be $U_X = \sum_{k=1}^{d_X} \frac{A_k}{2^k}$. Denote $\dot{A}_k = 2A_k - 1$. The σ-field generated by $\dot{A}_k$'s,

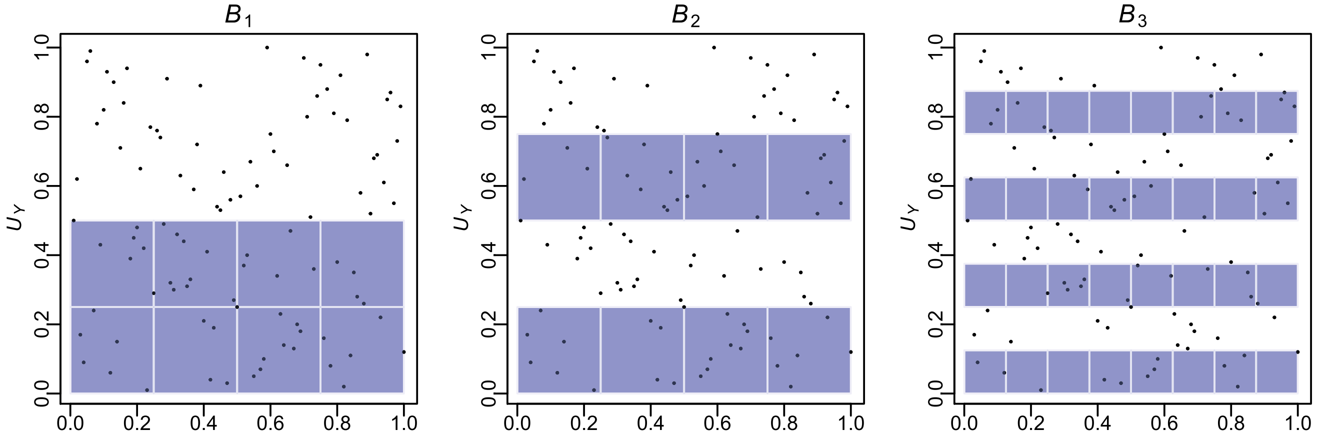

Figure 2. Binary variables B_1, B_2, B_3 from binary expansions of U_Y. Regions with $B_k = 1, k = 1, 2, 3$ are in white and regions with $B_k = 0, k = 1, 2, 3$ are in blue.

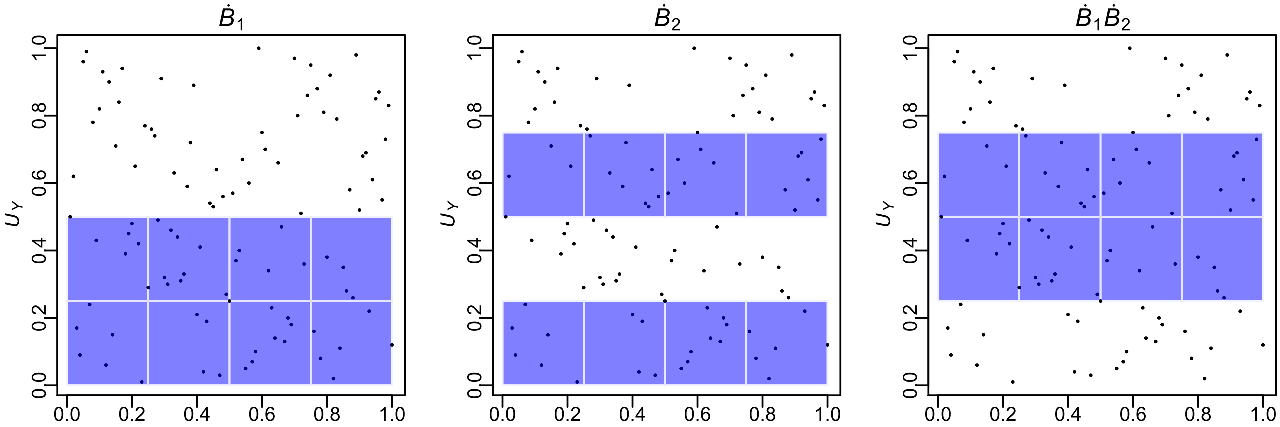

Figure 3. Basis binary variables in $\sigma(\dot{B}_1, \dot{B}_2)$, where U_Y is expanded up to the second order.

denoted by $\sigma(\dot{A}_1, \ldots, \dot{A}_{d_X})$, has a basis with $2^{d_X} - 1$ variables $W_{\dot{A}} = \{\dot{A}_1, \ldots, \dot{A}_{d_X}, \dot{A}_1\dot{A}_2, \ldots, \dot{A}_{d_X-1}\dot{A}_{d_X}, \ldots, \prod_{i=1}^{d_X} \dot{A}_i\}$. We shall refer to elements in $W_{\dot{A}}$ as patterns of X. After the construction of the patterns, the complicated effect of X on Y can be captured by logistic regression, which enjoys efficiency from the orthogonality of the patterns. We establish a set of $2^{d_Y} - 1$ penalized logistic regressions with the ℓ_1 penalty (Tibshirani 1996) on each resolution in $W_{\dot{B}}$, with all patterns in $W_{\dot{A}}$ as predictors. Denote the pattern vector corresponding to $W_{\dot{A}}$ by $\dot{\boldsymbol{A}}^F = (\dot{A}^F_{(1)}, \ldots, \dot{A}^F_{(2^{d_X}-1)})^T \triangleq (\dot{A}_1, \ldots, \dot{A}_{d_X}, \dot{A}_1\dot{A}_2, \ldots, \dot{A}_{d_X-1}\dot{A}_{d_X}, \ldots, \prod_{i=1}^{d_X} \dot{A}_i)^T$. Similarly, denote the resolution vector corresponding to $W_{\dot{B}}$ by $\dot{\boldsymbol{B}} = (\dot{B}_{(1)}, \ldots, \dot{B}_{(2^{d_Y}-1)})^T \triangleq (\dot{B}_1, \ldots, \dot{B}_{d_Y}, \dot{B}_1\dot{B}_2, \ldots, \dot{B}_{d_Y-1}\dot{B}_{d_Y}, \ldots, \prod_{i=1}^{d_Y} \dot{B}_i)^T$. Let $\{(x_i, y_i)\}_{i=1}^n$ be n independent observations of (X, Y). Denote $\dot{\boldsymbol{A}}_i^F = (\dot{A}^F_{(1),i}, \ldots, \dot{A}^F_{(2^{d_X}),i})^T$ and $\dot{\boldsymbol{B}}_i = (\dot{B}_{(1),i}, \ldots, \dot{B}_{(2^{d_Y}),i})^T$ as the ith pattern vector and the ith resolution vector obtained by binary expansions of the empirical CDF transformation of x_i and y_i, respectively. The mth logistic regression, $m = 1, \ldots, 2^{d_Y} - 1$, which models the effect of $\dot{\boldsymbol{A}}^F$ on $\dot{B}_{(m)}$, is established as

$$\log \frac{P(\dot{B}_{(m),i} = 1|\dot{\boldsymbol{A}}_i^F)}{P(\dot{B}_{(m),i} = -1|\dot{\boldsymbol{A}}_i^F)} = \dot{\boldsymbol{A}}_i^{F^T} \beta_m, \; m = 1, \ldots, 2^{d_Y} - 1, \quad (6)$$

where β_m is the coefficient vector. We employ a ℓ_1 regularization to give the estimator

$$\hat{\beta}_m = \arg\min_{\beta} \sum_{i=1}^{n} \log(1 + e^{-\dot{B}_{(m),i}\dot{A}_i^{F^T}\beta}) + \lambda||\beta||_1,$$
$$m = 1, \ldots, 2^{d_Y} - 1. \quad (7)$$

The conditional expectation of $\dot{B}_{(m)}$ given $\dot{\boldsymbol{A}}^F$, denoted by $e_m(\dot{\boldsymbol{A}}^F)$, can be estimated by $\hat{e}_m(\dot{\boldsymbol{A}}^F) = \frac{\exp(\dot{\boldsymbol{A}}^{F^T}\hat{\beta}_m)}{1+\exp(\dot{\boldsymbol{A}}^{F^T}\hat{\beta}_m)}$.

2.3. Binary Interaction Design: From Frequency Domain Back to Probability Domain

As a final step of estimating the distribution of Y, we aim to transform the conditional expectations of resolutions into the conditional cell probabilities of the corresponding histogram. First, we simplify the notation of the conditional expectation by using a d_Y-dimensional binary index. Namely, denote the conditional expectation $\mathbb{E}(\dot{B}_{k_1} \cdots \dot{B}_{k_p}|\dot{\boldsymbol{A}}^F)$, $p \in \{1, \ldots, d_Y\}$, $\{k_1, \ldots, k_p\} \subset \{1, \ldots, d_Y\}$, by $E_{\boldsymbol{b}}$, where $\mathbb{E}(\cdot)$ stands for the expectation, $\boldsymbol{b} = (b_1, \ldots, b_{d_Y})$ is a vector of length d_Y with value one at $k_1, \ldots, k_p$ and zero otherwise. Let $\boldsymbol{E}$ be the d_Y-dimensional conditional expectation vector whose entries are sorted in an ascending order according to the binary system. Hence, in some sense, $\boldsymbol{b}$ identifies the resolutions. Note that we set $E_{\mathbf{0}} = \mathbb{E}(1) = 1$ as the first entry. The $(\sum_{i=1}^{d_Y} b_i 2^{d_Y-i} + 1)$th entry is $E_{\boldsymbol{b}}$. For example, the expectation with $d_Y = 3$ is

$$\begin{aligned}\boldsymbol{E} &= (E_{000}, E_{001}, E_{010}, E_{011}, E_{100}, E_{101}, E_{110}, E_{111})^T,\\ &= (\mathbb{E}(1), \mathbb{E}(\dot{B}_3|\dot{\boldsymbol{A}}^F), \mathbb{E}(\dot{B}_2|\dot{\boldsymbol{A}}^F), \mathbb{E}(\dot{B}_2\dot{B}_3|\dot{\boldsymbol{A}}^F), \mathbb{E}(\dot{B}_1|\dot{\boldsymbol{A}}^F),\\ &\quad \mathbb{E}(\dot{B}_1\dot{B}_3|\dot{\boldsymbol{A}}^F), \mathbb{E}(\dot{B}_1\dot{B}_2|\dot{\boldsymbol{A}}^F), \mathbb{E}(\dot{B}_1\dot{B}_2\dot{B}_3|\dot{\boldsymbol{A}}^F))^T.\end{aligned}$$

We denote the conditional probabilities of the 2^{d_Y} cells in terms of the index $\boldsymbol{b} = (b_1, \ldots, b_{d_Y})$ as above. Define the conditional cell probability $p_{\boldsymbol{b}}$ as the conditional probability of $\dot{B}_k$'s taking values clarified by $\boldsymbol{b}$ given $\dot{\boldsymbol{A}}^F$, that is, $p_{\boldsymbol{b}} = p_{(b_1,\ldots,b_{d_Y})} = P\left(\dot{B}_1 = 2b_1 - 1, \ldots, \dot{B}_{d_Y} = 2b_{d_Y} - 1|\dot{\boldsymbol{A}}^F\right)$. As an example, for $d_Y = 3$, $p_{101} = P(\dot{B}_1 = 1, \dot{B}_2 = -1, \dot{B}_3 = 1|\dot{\boldsymbol{A}}^F)$. Let $\boldsymbol{p}$ be the d_Y-dimensional conditional probability vector of the cells whose entries are sorted by a descending order according to the binary system, that is, the $(2^{d_Y} - \sum_{i=1}^{d_Y} b_i 2^{d_Y-i})$th entry is $p_{\boldsymbol{b}}$. For example, the conditional probability vector of the cells with $d_Y = 3$ is $\boldsymbol{p} = (p_{111}, p_{110}, p_{101}, p_{100}, p_{011}, p_{010}, p_{001}, p_{000})^T$.

With the above notations, we establish the binary interaction design (BID) equation (Zhang 2019) to transform the expectations of resolutions into cell probabilities. The equation is established by the Sylvester's construction of Hadamard matrix $\boldsymbol{H} = \boldsymbol{H}_{2^{d_Y}}$ (Sylvester 1867).

Lemma 2 (BID equation). Let $\boldsymbol{E}$ be the conditional expectation vector of the resolutions from the binary expansion, and $\boldsymbol{p}$ be the conditional probability vector of the cells. Then

$$\boldsymbol{E} = \boldsymbol{H}\boldsymbol{p}, \quad (8)$$

where $\boldsymbol{H}$ is the Hadamard matrix (Sylvester 1867).

From the BID equation, the conditional probabilities of Y given X falling into each cell can be obtained by estimating the conditional expectation of resolutions. Denoting the estimator of $\boldsymbol{E}$ by $\hat{\boldsymbol{E}}$. In a common sense, $\boldsymbol{p}$ can be estimated by $\hat{\boldsymbol{p}} = \boldsymbol{H}^{-1}\hat{\boldsymbol{E}}$, since the Hadamard matrix $\boldsymbol{H}$ is invertible and $\boldsymbol{H}^{-1} = \frac{1}{2^{d_Y}}\boldsymbol{H}$. However, this $\hat{\boldsymbol{p}}$ may not be a probability measure. From the structure of the Hadamard matrix, the following lemma shows the summation of the cell probabilities is one.

Lemma 3. For any $\hat{\boldsymbol{E}}$, the estimation of $\boldsymbol{p}$ by $\hat{\boldsymbol{p}} = \boldsymbol{H}^{-1}\hat{\boldsymbol{E}}$ has a sum of entries of one.

Proof of Lemma 3. Denote $\hat{\boldsymbol{p}} = (\hat{p}_1, \dots, \hat{p}_{2^{d_Y}})^T$. Since $\hat{\boldsymbol{E}} = (1, \hat{e}_1(\dot{\boldsymbol{A}}^F), \dots, \hat{e}_{2^{d_Y}-1}(\dot{\boldsymbol{A}}^F))$, the summation of $\hat{\boldsymbol{p}}$ is

$$\sum_{i=1}^{2^{d_Y}} \hat{p}_i = \mathbf{1}_{2^{d_Y}}^T \hat{\boldsymbol{p}} = \frac{1}{2^{d_Y}} \mathbf{1}_{2^{d_Y}}^T \boldsymbol{H}^{-1}\hat{\boldsymbol{E}}$$
$$= \frac{1}{2^{d_Y}} (\ 2^{d_Y}, \quad 0, \quad \dots, \quad 0\ ) \hat{\boldsymbol{E}} = 1. \qquad \square$$

Note that we cannot guarantee $\hat{p}_i, i = 1, \dots, 2^{d_Y}$ to be positive. Instead of $\hat{\boldsymbol{p}} = \boldsymbol{H}^{-1}\hat{\boldsymbol{E}}$, we consider the following optimization problem to solve $\boldsymbol{p}$:

$$\begin{aligned} &\min ||\boldsymbol{H}\boldsymbol{p} - \hat{\boldsymbol{E}}||_1, \\ s.t. \quad & p_i \geqslant 0,\ i = 1, \dots, 2^{d_Y}, \\ & \sum_i p_i = 1. \end{aligned} \tag{9}$$

Since the cells can be viewed as the bins of the histogram of Y, we essentially estimate the distribution of Y as the resolutions are decomposed in an arbitrary delicate fashion. In practice, a finite d_Y is used, and we can smooth the histogram to approximate the distribution.

3. Multivariate Extensions

Now we extend our framework to the multivariate case. For a q-dimensional $X = (X_1, \dots, X_q)^T$, we perform a binary expansion to every marginal CDF-transformation variable, denoted by $U_{X_j}, j = 1, \dots, q$, up to the d_X-th order, and we have

$$U_{X_j} = \sum_{k=1}^{d_X} \frac{A_{jk}}{2^k}, j = 1, \dots, q. \tag{10}$$

Denote $\dot{A}_{jk} = 2A_{jk} - 1, j = 1, \dots, q, k = 1, \dots d_X$. The σ-field

$$\sigma(U_{X_1}, \dots, U_{X_q}) = \sigma(\dot{A}_{11}, \dots, \dot{A}_{1d_X}, \dots, \dot{A}_{q1}, \dots, \dot{A}_{qd_X})$$

is generated by the binary filtration of all q covariates. We can find a basis of this σ-field with totally $2^{qd_X} - 1$ variables. This basis set includes all possible patterns with respect to $X_j, j = 1, \dots, q$. These patterns can be divided into two groups. One group has terms involving binary variables from only one dimension of X, which capture the marginal patterns. For example, both the terms $\dot{A}_{11}$ and $\dot{A}_{11}\dot{A}_{12}$ are marginal patterns with respect to X_1. The second group has terms with binary variables from at least two covariates, which reflect interactions of the corresponding covariates. For example, $\dot{A}_{11}\dot{A}_{21}$ corresponds to the interaction of (X_1, X_2), and $\dot{A}_{11}\dot{A}_{21}\dot{A}_{31}$, $\dot{A}_{11}\dot{A}_{21}\dot{A}_{31}\dot{A}_{32}$ are terms with respect to the three-way interaction (X_1, X_2, X_3). Therefore, we refer to interaction terms as the patterns reflecting interaction of X, instead of product of some $\dot{A}_{jk}$'s. Note that the basis set considers until q-way interaction terms. However, three-way and higher-order interactions often contribute little to the model, and they are quite complex and difficult to interpret. Thus, we only consider the main effects and two-way interaction terms in the basis set. Including $2^{d_X} - 1$ main effect terms for each of the q explanatory variables, and $(2^{d_X} - 1)^2$ interaction terms of each pair of the explanatory variables, there are $L = q(2^{d_X} - 1) + C_q^2(2^{d_X} - 1)^2$ patterns in total.

To cope with high-dimensional data, some prescreening procedures can be performed to reduce the dimension of the patterns. We do not rashly reduce the maximum number of variables in a pattern term, since each of them includes information of some specific pattern, due to the orthogonal property of the binary expansion. Instead, an approach to pairwisely test the effect of each pattern on the response variable is reasonable. To this end, we modify the binary expansion testing (BET) method (Zhang 2019), which was originally developed to test independence of two variables, as a prescreening method for patterns. We also extend BET to a general version, which tests the independence of multiple variables and is used to prescreen interactions.

In the following, we first revisit the BET method, and extend it as a method of pattern prescreening in Section 3.1. In Section 3.2, we generalize BET to prescreen interactions by testing the independence of the response and the interaction patterns.

3.1. BET as a Prescreening Approach

BET is a nonparametric method of testing dependence between two continuous variables in a distribution-free setting. Hence, BET can be used on Y and X. With the binary expansion on U_Y and U_X, the interactions of the basis of the σ-field $\sigma(\dot{B}_1, \dots, \dot{B}_{d_Y})$ and $\sigma(\dot{A}_1, \dots, \dot{A}_{d_X})$ show all possible dependence patterns. Similar as the definition of $\boldsymbol{b}$ in Section 2.3, we use $\boldsymbol{a} = (a_1, \dots, a_{d_X})^T$ to identify the patterns of X. We denote the interaction pattern of $\boldsymbol{a}$ and $\boldsymbol{b}$ by $\boldsymbol{ab} := (a_1, \dots, a_{d_X}, b_1, \dots, b_{d_Y})^T$. The interaction pattern $\boldsymbol{ab}$ can partition the unit square $[0, 1]^2$ with half positive regions and half negative regions. The difference of the counts in the two regions reflects whether Y and X are independent in terms of the particular interaction pattern. When U_Y and U_X are independent, the counts of the observations in the positive and negative regions should be similar. When they are not independent, there will be significant difference of counts. Denote $S_{\boldsymbol{ab}}$ as the difference of the counts with respect to the interaction pattern $\boldsymbol{ab}$. Zhang (2019) gave the result of the distribution of $S_{\boldsymbol{ab}}$ in the following lemma.

Lemma 4.

1. When marginal distributions are known, U_Y and U_X are independent if and only if

$$\frac{S_{\boldsymbol{ab}} + n}{2} \sim \text{Binomial}(n, \frac{1}{2}),\ \boldsymbol{a} \neq \mathbf{0}, \boldsymbol{b} \neq \mathbf{0}.$$

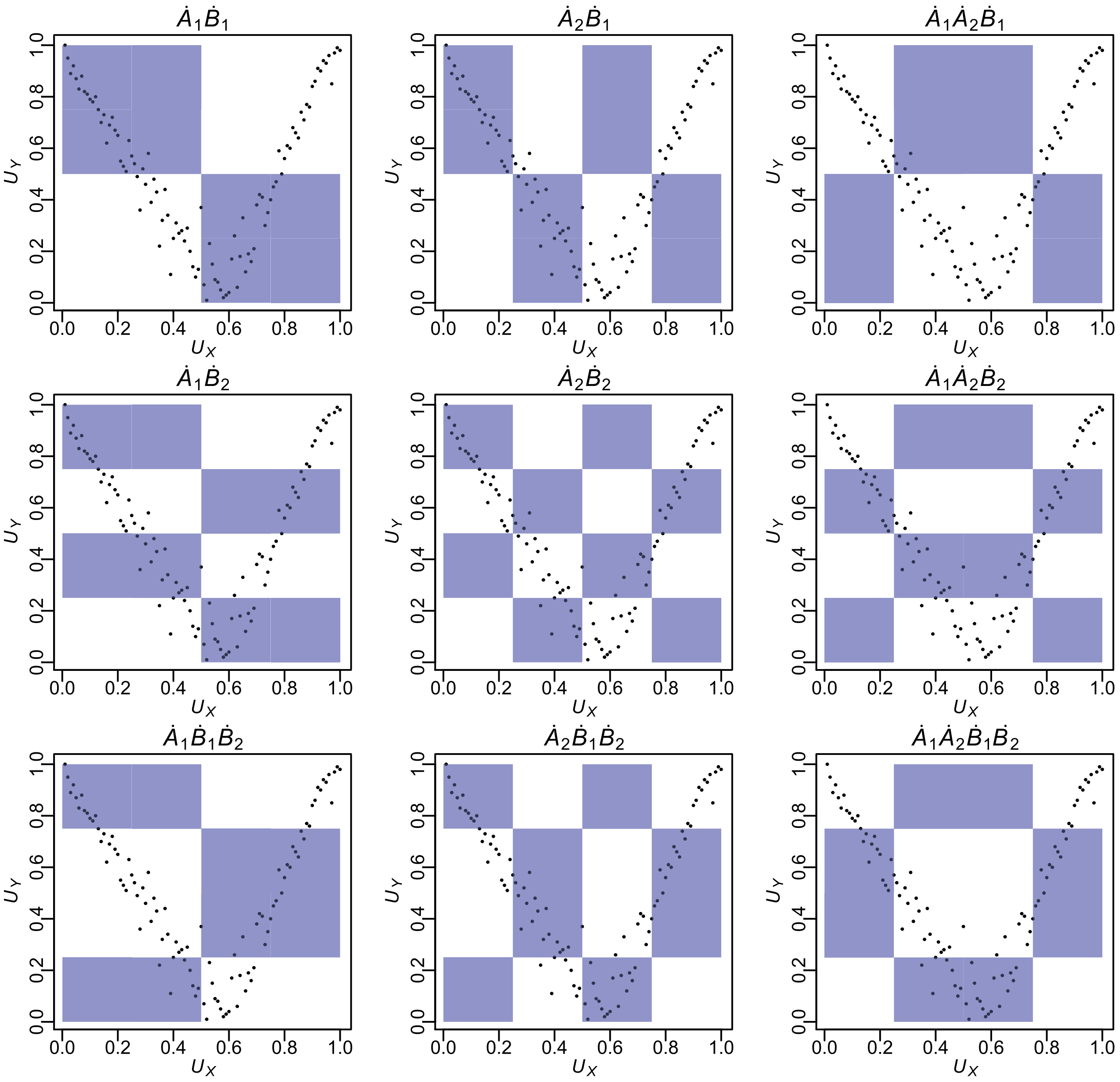

Figure 4. The dependence of Y on X is from the model $Y = X^2 + \varepsilon$, where $X \sim$ Uniform$(-2, 2)$ and $\varepsilon \sim N(0, 0.25)$. BET detects the dependence through the nine patterns. The pattern $\dot{A}_1\dot{A}_2\dot{B}_1$ shows the most obvious difference of counts of blue and white regions.

2. When marginal distributions are unknown, U_Y and U_X are estimated by the empirical CDF transformations $\widehat{U}_Y$ and $\widehat{U}_X$, respectively, then $\widehat{U}_Y$ and $\widehat{U}_X$ are independent if and only if

$$\frac{\widehat{S}_{\boldsymbol{ab}} + n}{4} \sim \text{Hypergeometric}(n, \frac{n}{2}, \frac{n}{2}), \ \boldsymbol{a} \neq \boldsymbol{0}, \boldsymbol{b} \neq \boldsymbol{0},$$

where $\widehat{S}_{\boldsymbol{ab}}$ denotes the difference of the counts with respect to the interaction pattern $\boldsymbol{ab}$ according to $\widehat{U}_Y$ and $\widehat{U}_X$.

In this way, BET decomposes the information of the relationship between Y and X into interaction patterns. Figure 4 shows all the nine interaction patterns with depth $d_X = 2$ and $d_Y = 2$. An obvious dependence pattern is $\dot{A}_1\dot{A}_2\dot{B}_1$, which includes most points in the white region.

With d_X and d_Y large enough, BET can detect arbitrarily complicated dependence. BET also helps indicate the pattern of dependence, since the significant patterns from BET imply how Y depends on X. This inspires us to focus on the detection of the significant patterns and regard BET as a pattern-screening approach. Performing BET pairwisely on $\{(Y, X_j), j = 1, \dots q\}$, one can reduce all the patterns on X_j to only those dependent ones. We regard the patterns of X that are detected to be dependent with at least one resolution of Y as relevant variables in the penalized logistic regressions. Namely, denoting the pattern vector of X_j by $\dot{\boldsymbol{A}}^j = (\dot{A}^j_{(1)}, \dots, \dot{A}^j_{(2^{d_X}-1)})^T \triangleq (\dot{A}^j_1, \dots, \dot{A}^j_{d_X}, \dot{A}^j_1\dot{A}^j_2, \dots, \dot{A}^j_{d_X-1}\dot{A}^j_{d_X}, \dots, \prod_{i=1}^{d_X} \dot{A}^j_i)^T$, we obtain the sets of relevant patterns $R_j^{\text{main}} := \{\dot{A}^j_{(l)} : \exists \dot{B}_{(m)} s.t. (\dot{B}_{(m)}, \dot{A}^j_{(l)})$ *is* dependent, $l = 1, \dots, 2^{d_X} - 1\}$, $j = 1, \dots, q$. Note that we consider the patterns in $\cup_{j=1}^{q} R_j^{\text{main}}$ as predictors in regressions for all resolutions of Y, rather than only in the regression for the particular $\dot{B}_{(m)}$ such that $(\dot{B}_{(m)}, \dot{A}^j_{(l)})$ is dependent. This can

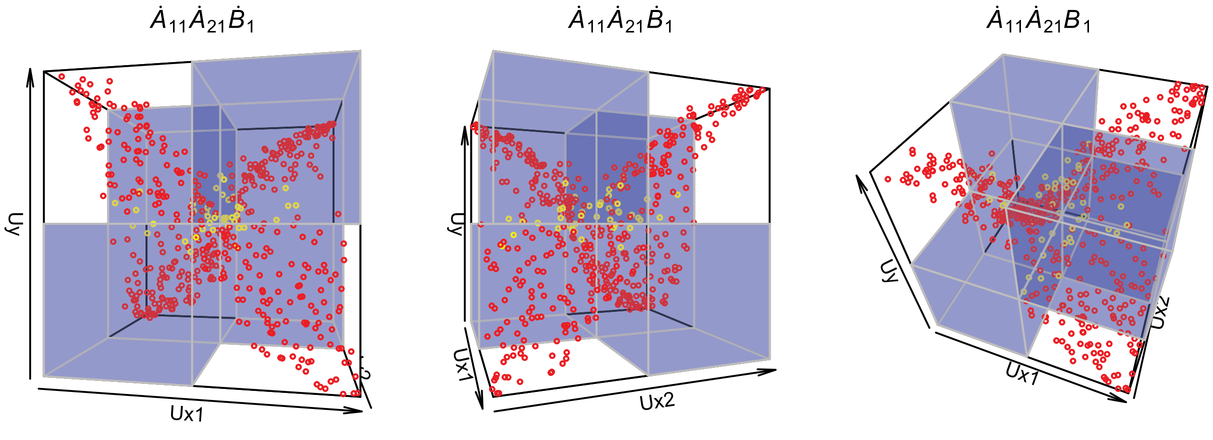

Figure 5. The dependence of Y and X_1, X_2 is from the model $Y = X_1X_2 + \varepsilon$, where $X_1, X_2 \overset{\text{iid}}{\sim} \text{Uniform}(-2, 2)$ and $\varepsilon \sim N(0, 0.25)$. The generalized BET detects the pattern $\dot{A}_{11}\dot{A}_{21}\dot{B}_1$ with the most obvious difference in counts of the positive region (white region with red points) and the negative region (blue region with yellow points).

help to avoid false negatives. False positives can be controlled by the lasso shrinkage.

3.2. A Generalized BET as an Interaction Prescreening Method

The pairwise BET procedure selects dependent marginal patterns. Furthermore, aiming to select interaction patterns, we first generalize the original BET to test independence of Y and the joint distribution of X_i and $X_j, i, j = 1, \dots, q$. With marginal binary expansions on U_Y, U_{X_i}, U_{X_j}, respectively, the σ-field generated by U_Y, U_{X_i} and U_{X_j} are $\sigma(U_Y) = \sigma(\dot{B}_1, \dots, \dot{B}_{d_Y})$ and $\sigma(U_{X_i}, U_{X_j}) = \sigma(\dot{A}_{i1}, \dots, \dot{A}_{id_X}, \dot{A}_{j1}, \dots, \dot{A}_{jd_X})$. We aim to test all possible dependence patterns from each pair of the two σ-fields. Denote the pattern of X_j, $j = 1, \dots, q$, by $\boldsymbol{a}_j = (a_{j1}, \dots, a_{jd_X})^T$, and the three-way interaction pattern of $\boldsymbol{a}_i$, $\boldsymbol{a}_j$ and $\boldsymbol{b}$ by $\boldsymbol{a}_i\boldsymbol{a}_j\boldsymbol{b} = (a_{i1}, \dots, a_{id_X}, a_{j1}, \dots, a_{jd_X}, b_1, \dots, b_{d_Y})^T$. Similar to the idea of the original BET, $\boldsymbol{a}_i\boldsymbol{a}_j\boldsymbol{b}$ can be viewed as a partition of the cube $[0, 1]^3$ with half positive and half negative regions. One can test the dependence of Y and the joint X_i, X_j by the different counts of the two regions $S_{\boldsymbol{a}_i\boldsymbol{a}_j\boldsymbol{b}}$. Figure 5 shows three aspects of a significant interaction pattern. Hence, we use the generalized BET to prescreen the interaction predictors in the regressions. We perform the generalized BET pairwise on $\{(Y, X_i, X_j), j = 1, \dots, q, i = 1, \dots, j-1\}$ and obtain the sets of significant interactions

$$R_{ij}^{\text{interaction}} = \{(\dot{A}^i_{(l_1)}, \dot{A}^j_{(l_2)}) : \exists \dot{B}_{(q)} \; s.t. \; (\dot{B}_{(m)}, \dot{A}^i_{(l_1)}, \dot{A}^j_{(l_2)}) \text{ is dependent, } l_1, l_2 = 1, \dots, 2^{d_X} - 1\},$$
$$j = 1, \dots, q, \; i = 1, \dots, j.$$

With the two prescreening procedures, eventually, the predictor set is

$$R = \left(\cup_{i=1}^{q} R_i^{\text{main}}\right) \cup \left(\cup_{i<j} R_{ij}^{\text{interaction}}\right). \tag{11}$$

We refer to the prescreening based on BET and the generalized BET as the BET screening. Algorithm 1 gives the procedure of resolution-wise regression, including the prescreening and the framework of the estimation.

Algorithm 1

Step 1. Consider the binary expansions of U_Y and $U_{X_j}, j = 1, \dots, q$ as in (4) and (10). List the binary variables in the resolutions of Y in $W_{\dot{B}}$ as in (5). List the binary variables in the patterns of X in $W_{\dot{A}}$ similarly.

Step 2. Prescreen the main effects and the interactions by BET screening, respectively, and obtain the relevant predictors as in (11).

Step 3. Perform a set of $2^{d^Y} - 1$ penalized logistic regressions as in (7) and obtain the estimated expectation vector $\hat{\boldsymbol{E}}$.

Step 4. Obtain the cell probability vector by the optimization problem (9).

4. Theoretical Studies

In this section, we first show that the BET screening is a sure independence screening approach (Fan and Song 2010), which reduces the number of patterns L from exponential growth to $O(n)$. The consistency result of the estimated cell probabilities is also established with the random design and a fixed d_Y. We allow the dimension q and d_X to grow with n.

For the mth logistic regression, assume that the binary data $\{Z_{m,i}\}_{i=1}^n = \{(\dot{\boldsymbol{A}}_i, \dot{B}_{(m),i})\}_{i=1}^n$ from the marginal empirical CDF transformation observations $\{(x_i, y_i)\}_{i=1}^n$ are iid copies of $(\dot{\boldsymbol{A}}, \dot{B}_{(m)})$, where $\dot{\boldsymbol{A}}_i = (\dot{A}_{(1),i}, \dots, \dot{A}_{(L),i})^T$ is the ith sample of the L-dimensional binary random vector $\dot{\boldsymbol{A}} = (\dot{A}_{(1)}, \dots, \dot{A}_{(L)})^T$, and $\dot{B}_{(m),i}$ is the ith sample of the binary response $\dot{B}_{(m)}$. Denote $\tilde{\boldsymbol{A}}_i := (\tilde{A}_{(1),i}, \dots, \tilde{A}_{(L),i})^T$, $i = 1, \dots, n$, as the samples of the covariates $\tilde{\boldsymbol{A}} = (\tilde{A}_{(1)}, \dots, \tilde{A}_{(L)})$ that are standardized to have mean zero and standard deviation one for each covariate. We have $\tilde{\boldsymbol{A}}_i = \dot{\boldsymbol{A}}_i/\sqrt{n}$. Denote $\tilde{B}_{(m),i}$ as the ith sample of $\tilde{B}_{(m)}$ taking values from $\{0, 1\}$. We have $\tilde{B}_{(m)} = (\dot{B}_{(m)} + 1)/2$. The maximum marginal likelihood estimator (MMLE) $\tilde{\beta}_{m,j}$ for the logistic regression (6), which is a special case of the models in Fan and Song (2010), is defined as the minimizer of the negative log-likelihood of the component-wise regression,

$$\tilde{\beta}_{m,j} = \arg\min_{\beta_{m,j}} \frac{1}{n} \sum_{i=1}^{n} -\tilde{B}_{(m),i}\tilde{A}_{(j),i}\beta_{m,j} + \log(1 + e^{\tilde{A}_{(j),i}\beta_{m,j}}), \; j = 1, \dots, L. \tag{12}$$

We correspondingly define the population version of the MMLE by

$$\beta^M_{m,j} = \arg\min_{\beta_{m,j}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[-\tilde{B}_{(m),i}\tilde{A}_{(j),i}\beta_{m,j} + \log(1+e^{\tilde{A}_{(j),i}\beta_{m,j}})],\ j=1,\ldots,L.$$

Denote the true regression coefficient vector by $\beta^0_m := (\beta^0_{m,1},\ldots,\beta^0_{m,L})^T$. Let $\mathcal{M}^*_m := \{1 \leqslant j \leqslant L : \beta^0_{m,j} \neq 0\}$ be the true index set of nonzero coefficients. We remark here that our overall goal of the analysis is prediction of the response rather than inference of slopes. Therefore, although when d_Y and d_X are large, and the overall parameterization might become unidentifiable, it will not harm the prediction results, as studied in Greenshtein and Ritov (2004).

We now provide the theoretical justifications of our method.

Assumption 1. $\left|\mathrm{cov}\left(\frac{e^{\tilde{A}^T\beta^0_m}}{1+e^{\tilde{A}^T\beta^0_m}}, \tilde{A}_{(j)}\right)\right| \geqslant c_{1,m}n^{-\kappa_m}$ for $j \in \mathcal{M}^*_m$ with constants $c_{1,m} > 0$ and $0 < \kappa_m < 1/2$.

Assumption 1 is analogous to Condition E of Fan and Song (2010). It ensures that the marginal signals are stronger than the stochastic noise. Within the selected set R, denote $\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}$ as the pattern corresponding to $\dot{A}_{(j)}$ and $\dot{B}_{(m)}$, and let the index set of selected variables using BET screening be $\mathcal{M}_{m,\delta_{n,m}} := \{1 \leqslant j \leqslant L : S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} \geqslant \delta_{n,m}\}$, for some threshold $\delta_{n,m}$. The following theorem shows that BET screening possesses the sure independence screening property.

Theorem 1. For any $c_{2,m} > 0$, there exists a positive constant $c_{3,m}$ such that

$$P(\max_{1\leqslant j\leqslant L} |\tilde{\beta}_{m,j} - \beta^M_{m,j}| \geqslant c_{2,m}n^{-\kappa_m}) \leqslant L\{\exp(-c_{3,m}n^{1-2\kappa_m}/(k_{n,m}K_{n,m})^2) + nh_{1,m}\exp(-h_{0,m}K^{\alpha_m}_{n,m})\},$$

where $k_{n,m}$, $K_{n,m}$, $h_{0,m}$, $h_{1,m}$, α_m are some positive constants. If, in addition, Assumption 1 holds, the BET screening possesses a sure independence screening property. By taking $\delta_{n,m} = O(n^{\frac{1}{2}-\kappa_m})$, we have

$$P(\mathcal{M}^*_m \subset \mathcal{M}_{m,\delta_{n,m}}) \geqslant 1 - s_m\{\exp(-c_{3,m}n^{1-2\kappa_m}/(k_{n,m}K_{n,m})^2) + nh_{1,m}\exp(-h_{0,m}K^{\alpha_m}_{n,m})\},$$

where $s_m := |\mathcal{M}^*_m|$, the number of nonsparse elements.

Assumption 2. The variance $\mathrm{var}(\tilde{\boldsymbol{A}}^T\beta^0_m)$ is bounded from above and below.

Assumption 2 is analogous to Condition F of Fan and Song (2010). The following theorem shows that the BET screening can reduce the dimension from $O(q^2 4^{d_X})$ to $O(n^{2\kappa_m})$.

Theorem 2. Under Assumption 2, we have for any $\delta_{n,m} = O(n^{\frac{1}{2}-\kappa_m})$, and the same constants $c_{3,m}$, $k_{n,m}$, $K_{n,m}$, $h_{0,m}$, $h_{1,m}$, α_m as in Theorem 1 such that

$$P(|\mathcal{M}_{m,\delta_{n,m}}| \leqslant O(n^{2\kappa_m})) \geqslant 1 - L\{\exp(-c_{3,m}n^{1-2\kappa_m}/(k_{n,m}K_{n,m})^2) + nh_{1,m}\exp(-h_{0,m}K^{\alpha_m}_{n,m})\}.$$

Here, we briefly describe the results, whose details are given in the Appendix. Let $r := \max_{1\leqslant m\leqslant 2^{d_Y}-1} |\mathcal{M}_{m,\delta_{n,m}}|$ and $\dot{\boldsymbol{A}}^S := (\dot{A}^S_{(1)},\ldots,\dot{A}^S_{(r)})^T$ is the predictor vector including the selected r patterns. Denote the true coefficient vector of β_m by β^0_m. The estimate of β_m is

$$\hat{\beta}_m = \arg\min_{\beta} \sum_{i=1}^{n} \log(1+e^{-\dot{B}_{(m),i}f_m(\dot{\boldsymbol{A}}^S_i)}) + \lambda_m||\beta||_1\},\ m=1,\ldots,2^{d_Y}-1,$$

where $||\cdot||_1$ is the ℓ_1-norm, λ_m is a tuning parameter, $\dot{\boldsymbol{A}}^S_i$ is the binary expansion corresponding to $\dot{\boldsymbol{A}}^S$ for the ith observation, and f_m is the mth logistic regression function, $m = 1,\ldots,2^{d_Y}-1$. Let f^0_m be the true function between $\dot{B}_{(m)}$ and $\dot{\boldsymbol{A}}^S$. Denote the index set of nonzero coefficients by $S^0_m := \{j : \beta^0_{m,j} \neq 0\}$, and the cardinality of S^0_m by $s_m := |S^0_m|$.

According to the BID equation, we estimate $\boldsymbol{p}$ by solving the optimization (9). From the optimization, $\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}}$ is an approximation of $\hat{e}_m$, where $\boldsymbol{H}_{m+1}$ is the $(m+1)$th row of $\boldsymbol{H}$, since $\hat{e}_m$ is the $(m+1)$th entry of $\hat{\boldsymbol{E}}$. Hence, $g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})$ is the estimated mth regression function corresponding to $\hat{\boldsymbol{p}}$. The following theorem gives the consistency of cell probability vector $\hat{\boldsymbol{p}}$ in terms of excess risk of $g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})$.

Theorem 3. Assume Assumptions 1 and 2, and 3–5 given in the Appendix hold, where Assumption 3 in the Appendix holds with the set S^0_m. For the logistic regression with covariates corresponding to the BET screening set $\mathcal{M}_{m,\delta_{n,m}}$, suppose that λ_m satisfies $\lambda_m \geqslant 8\lambda^0_m$. Then on the set $\mathcal{T}_m$, we have,

$$\mathcal{E}(g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})) + \lambda_m||\hat{\beta}_m - \beta^0_m||_1 \leqslant 6\mathcal{E}(f^0_m) + \frac{16\lambda^2_m s_m}{c_m\phi^2_m} + \frac{32\lambda K s 2^{d_Y}}{c\phi^2},$$

where $K > 0$ is a constant, $\lambda = \max_{1\leqslant m\leqslant 2^{d_Y}-1} \lambda_m$, $s = \max_{1\leqslant m\leqslant 2^{d_Y}-1} s_m$, $c = \min_{1\leqslant m\leqslant 2^{d_Y}-1} c_m$, ϕ^2_m is a compatibility constant, and $\phi = \min_{1\leqslant m\leqslant 2^{d_Y}-1} \phi_m$.

5. Simulation Studies

In this section, we perform simulations to show the performance of resolution-wise regression approach. We compare our method with the following four methods:

1. Naive method, which first finds a small neighborhood of each test sample in the training set, where $||X - X_{\text{new}}||_2$ is bounded by a constant, and predicts the distribution of $Y|X_{\text{new}}$ by the kernel density estimation of the responses in this neighborhood.
2. SSANOVA, which fits a cubic spline with all main effects and interaction effects. Its prediction distribution is $Y|X_{\text{new}} \sim N(\hat{Y}_{\text{ssanova}}|X_{\text{new}}, \hat{\sigma}^2+\mathrm{var}(\hat{Y}_{\text{ssanova}}|X_{\text{new}}))$, where $\hat{Y}_{\text{sanova}}|X_{\text{new}}$ is the SSANOVA estimation of Y given a new X_{new}, and $\hat{\sigma}^2$ is the estimated variance of the random error.

3. Random Forest, which fits a multitude of regression trees and then averages the predictions. Its prediction distribution is $Y|X_{\text{new}} \sim N(\hat{Y}_{rf}|X_{\text{new}}, \hat{\sigma}_s{}^2)$, where $\hat{Y}_{rf}|X_{\text{new}}$ is the estimation of Y from random forest given a new X_{new}, and $\hat{\sigma}_s{}^2$ is the standard error.
4. Regression mixture model (only for Example 2), which identifies the subgroups of dataset and fits multiple linear regression models. Its prediction distribution is $Y|X_{\text{new}} \sim N(\hat{Y}_{\text{mixreg}}|X_{\text{new}}, \hat{\sigma}_s{}^2)$, where $\hat{Y}_{\text{mixreg}}|X_{\text{new}}$ is the estimation of Y from the regression mixture model given a new X_{new} which is randomly assigned into subgroups with the weights derived from training data, and $\hat{\sigma}_s^2$ is the standard error.

We study the following four examples with 1024 samples for both training and testing sets.

Example 1 (Crossing lines). The predictor $x_i \overset{\text{iid}}{\sim} U(-10, 10)$, $i = 1, \dots, n$. For the example with one cross on the plane, the response y_i is generated by $y_i = x_i I(g_i = 0) - x_i I(g_i = 1) + \varepsilon_i$, where the error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.5)$, $i = 1, \dots, n$, and $I(\cdot)$ is the indicator function with $g_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$, $i = 1, \dots, n$. For the example with multiple crosses, the response $y_i = \big(x_i I(g_{1i} = 1) - x_i I(g_{1i} = 2) + (x_i - 10) I(g_{1i} = 3) + (-x_i + 10) I(g_{1i} = 4)\big) I(x_i \geq 0) + \big(x_i I(g_{2i} = 1) - x_i I(g_{2i} = 2) + (-x_i - 10) I(g_{2i} = 3) + (x_i + 10) I(g_{2i} = 4)\big) I(x_i < 0) + \varepsilon_i$, where the error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.5)$, $i = 1, \dots, n$, and $I(\cdot)$ is the indicator function with $g_{ki} \overset{\text{iid}}{\sim} \text{Multi-Bern}(\{1, 2, 3, 4\}, (1/4, 1/4, 1/4, 1/4))$, $i = 1, \dots, n$, $k = 1, 2$.

Example 2 (A mixture of linear and quadratic effects). The predictor vector $(x_{i1}, \dots, x_{iq})^T$ is generated by $x_{ij} \overset{\text{iid}}{\sim} U(-2, 2)$, $i = 1, \dots, n$, $j = 1, \dots, q$, with $q = 1, 5, 10$. The response y_i is generated by $y_i = x_{i1} I(g_i = 0) + x_{i1}^2 I(g_i = 1) + \varepsilon_i$, which depends on only the first variable x_{i1} and other variables are regarded as noise. The error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.05)$, $i = 1, \dots, n$, and $I(\cdot)$ is the indicator function with $g_i \overset{\text{iid}}{\sim} \text{Bernoulli}(1/2)$, $i = 1, \dots, n$.

Example 3 (Circular and spherical implicit functional relationship). The predictor vector $(x_{i1}, \dots, x_{iq})^T$ has $q = 5$. For the circle example, the predictors and the responses are generated from the polar coordinates $x_{i1} = \sin(\theta_i)$, $y_i = \cos(\theta_i) + \varepsilon_i$, where the latent variable $\theta_i \overset{\text{iid}}{\sim} U(0, 2\pi)$, $i = 1, \dots, n$, and the error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.05)$, $i = 1, \dots, n$. The noise variables $(x_{i2}, \dots, x_{iq})^T$ are generated by $x_{ij} \overset{\text{iid}}{\sim} U(-1, 1)$, $i = 1, \dots, n$, $j = 2, \dots, q$. For the sphere example, where the latent variables $\theta_i \overset{\text{iid}}{\sim} U(0, \pi)$, the predictors and the responses are generated from $x_{i1} = \sin(\theta_i)\cos(\phi_i)$, $x_{i2} = \sin(\theta_i)\sin(\phi_i)$, $y_i = \cos(\theta_i) + \varepsilon_i$, where $\phi_i \overset{\text{iid}}{\sim} U(0, 2\pi)$, $i = 1, \dots, n$, and the error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 0.05)$, $i = 1, \dots, n$. The noise variables $(x_{i3}, \dots, x_{iq})^T$ are generated by $x_{ij} \overset{\text{iid}}{\sim} U(-1, 1)$, $i = 1, \dots, n$, $j = 3, \dots, q$.

Example 4 (Heterogeneous mean vs. heteroscedastic error). The predictor $x_i \overset{\text{iid}}{\sim} U(-2, 2)$, $i = 1, \dots, n$. The response y_i is generated by $y_i = (x_i + 2 + x_i\varepsilon_i) I(g_i = 0) + (x_{i1}^2 + \varepsilon_i) I(g_i = 1)$, where the error $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$, $i = 1, \dots, n$, $\sigma^2 = 0.05,\ 0.25,\ 0.5$.

In each example, BET with depth 5 and a threshold of $\sqrt{\frac{2\log p}{n}}$ for symmetry statistics, where p is the total number of interactions and n is the sample size, is performed for main effects screening, while generalized BET with depth 4 and the same threshold is performed for interaction effects screening. For the selection of depth, with a small depth 3, BET reaches a high power (Zhang, Zhao and Zhou 2021). We perform simulation with different depths and the results are reported in the Appendix. We pick a depth 5 and 4, which is high enough. Two types of smoothing approaches are considered: fixed smoothing parameter ("Fixed smoothness"), and tuning the smoothing parameter by cross-validation ("CV").

We repeat the simulation 100 times for each example. To measure the test error, we calculate the differences of prediction distributions from different methods and the underlying true distributions, the following distance measures are used: (a) Kolmogorov-Smirnov statistic $D_{KS}(P, Q) = \sup_x |P(x) - Q(x)|$, (b) Kullback-Leibler divergence $D_{KL} = \int_{-\infty}^{\infty} p(x) \log(\frac{p(x)}{q(x)}) dx$, (c) L_1 distance $D_{L_1} = \int_{-\infty}^{\infty} |p(x) - q(x)| dx$, where P and Q are two distributions with corresponding PDFs $p(\cdot)$ and $q(\cdot)$, respectively.

Here we display the results of Example 1–4, which are heterogeneous data, and the result from a case of a nonlinear functional relationship is in supplementary materials. Tables 1–4 list the results of testing errors for the three simulation examples. Figures 6–9 show the heatmaps of the prediction distributions of all test data, where the x axis is the involved variable of the predictors. We discard the heatmap of the spherical case, since it has two involved variables and cannot be shown explicitly in a heatmap.

The results indicate the best performance of resolution-wise regression. For Example 1, resolution-wise regression especially with cross-validation smoothness can identify the subgroups, while the regression mixture model does not perform well as the number of subgroups increases. The naive method has a good performance since the dependence is linear. For

Table 1. Comparison of average test errors (and corresponding standard errors in parentheses) for Example 1 with respect to one or multiple crosses and three distance measures. Bold numbers represent the smallest error of each case.

Example	Measure	Naive	Mixreg	SSANOVA	Random Forest	Fixed smoothness	CV
one cross	KS	0.151	0.351	0.294	0.377	0.165	**0.129**
		(0.015)	(0.010)	(0.016)	(0.017)	(0.009)	(0.005)
	KL	0.362	1.396	1.031	1.835	0.386	**0.265**
		(0.032)	(0.067)	(0.062)	(0.156)	(0.025)	(0.014)
	L_1	0.639	1.086	1.181	1.120	0.717	**0.364**
		(0.034)	(0.063)	(0.073)	(0.069)	(0.035)	(0.019)
multiple	KS	0.188	0.405	0.229	0.394	0.165	**0.145**
		(0.007)	(0.026)	(0.016)	(0.020)	(0.010)	(0.008)
	KL	0.319	2.511	0.548	1.615	0.338	**0.257**
		(0.016)	(0.128)	(0.036)	(0.084)	(0.024)	(0.016)
	L_1	0.653	1.212	0.850	1.054	0.658	**0.397**
		(0.033)	(0.052)	(0.042)	(0.057)	(0.034)	(0.019)

NOTE: The results for naive method, mixture of regression, SSANOVA, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV are listed in the columns from left to right, respectively.

Table 2. Comparison of average test errors (and corresponding standard errors in parentheses) for Example 2 with respect to dimension $q = 1, 5, 10$ and three distance measures. Bold numbers represent the smallest error of each case.

Example	Measure	Naive	SSANOVA	Random Forest	Fixed smoothness	CV
$q = 1$	KS	**0.103**	0.215	0.256	0.169	0.167
		(0.006)	(0.005)	(0.008)	(0.013)	(0.014)
	KL	0.261	0.605	0.787	**0.089**	0.693
		(0.010)	(0.021)	(0.049)	(0.008)	(0.037)
	L_1	0.383	0.765	0.716	**0.283**	0.517
		(0.018)	(0.033)	(0.035)	(0.015)	(0.027)
$q = 5$	KS	0.345	0.219	0.222	0.180	**0.179**
		(0.019)	(0.016)	(0.012)	(0.015)	(0.014)
	KL	0.701	0.589	0.730	**0.146**	0.532
		(0.034)	(0.025)	(0.037)	(0.011)	(0.028)
	L_1	0.937	0.774	0.693	**0.348**	0.506
		(0.039)	(0.035)	(0.028)	(0.015)	(0.022)
$q = 10$	KS	0.344	0.210	0.213	**0.169**	0.183
		(0.006)	(0.006)	(0.015)	(0.018)	(0.018)
	KL	0.767	0.589	0.666	**0.185**	0.572
		(0.033)	(0.024)	(0.050)	(0.022)	(0.035)
	L_1	0.952	0.735	0.693	**0.346**	0.498
		(0.038)	(0.032)	(0.035)	(0.015)	(0.026)

NOTE: The results for naive method, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV are listed in the columns from left to right, respectively.

Table 3. Comparison of average test errors (and corresponding standard errors in parentheses) for Example 3 with respect to circular and spherical implicit functional relationship and three distance measures. Bold numbers represent the smallest error of each case.

Example	Measure	Naive	SSANOVA	Random Forest	Fixed smoothness	CV
Circle	KS	0.175	0.201	0.398	0.172	**0.156**
		(0.010)	(0.014)	(0.030)	(0.010)	(0.008)
	KL	0.378	0.436	4.180	**0.264**	0.404
		(0.019)	(0.032)	(0.322)	(0.008)	(0.027)
	L_1	0.677	0.761	1.384	**0.515**	0.535
		(0.039)	(0.034)	(0.108)	(0.022)	(0.025)
Sphere	KS	0.170	0.184	0.413	0.183	**0.161**
		(0.012)	(0.010)	(0.037)	(0.009)	(0.006)
	KL	0.369	0.411	4.553	**0.310**	0.339
		(0.015)	(0.021)	(0.355)	(0.009)	(0.016)
	L_1	0.664	0.735	1.439	0.580	**0.570**
		(0.026)	(0.038)	(0.080)	(0.024)	(0.025)

NOTE: The results for naive method, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV are listed in the columns from left to right, respectively.

Table 4. Comparison of average test errors (and corresponding standard errors in parentheses) for Example 4 with respect to different variances of random errors and three distance measure. Bold numbers represent the smallest error of each case.

Example	Measure	Naive	SSANOVA	Random Forest	Fixed smoothness	CV
$\sigma^2 = 0.05$	KS	0.192	0.189	0.243	**0.153**	0.154
		(0.008)	(0.009)	(0.016)	(0.005)	(0.005)
	KL	0.435	0.417	0.512	**0.229**	0.743
		(0.022)	(0.026)	(0.030)	(0.012)	(0.034)
	L_1	0.656	0.622	0.652	**0.424**	0.569
		(0.034)	(0.035)	(0.033)	(0.022)	(0.024)
$\sigma^2 = 0.25$	KS	0.156	0.131	0.239	**0.123**	**0.123**
		(0.006)	(0.006)	(0.014)	(0.009)	(0.010)
	KL	0.258	0.245	0.430	**0.166**	0.330
		(0.016)	(0.014)	(0.026)	(0.009)	(0.020)
	L_1	0.491	0.452	0.571	**0.346**	0.400
		(0.025)	(0.030)	(0.033)	(0.017)	(0.023)
$\sigma^2 = 0.5$	KS	0.139	0.121	0.236	**0.098**	0.110
		(0.008)	(0.007)	(0.018)	(0.007)	(0.009)
	KL	0.201	0.208	0.421	**0.142**	0.167
		(0.013)	(0.017)	(0.035)	(0.009)	(0.012)
	L_1	0.422	0.408	0.554	**0.296**	0.308
		(0.035)	(0.024)	(0.038)	(0.019)	(0.020)

NOTE: The results for naive method, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV are listed in the columns from left to right, respectively.

Example 2, resolution-wise regression can predict the probabilities around the two subgroups, thus, has the best performance. SSANOVA does not perform well since it cannot recognize the subgroups. The naive method performs well only in the low-dimensional case, because in the high-dimensional case, it is difficult to find a small neighborhood with substantial train data. Random forest does not perform well because it averages the predictions from multiple regression trees and mixes the two subgroups up. Resolution-wise regression performs well in the high-dimensional case, and the distance to the true distribution only increases slightly with the effect of noise variables. For Example 3, resolution-wise regression performs the best, while the naive method has poor performance due to the dimension issue; SSANOVA fails to capture the relationship, since it cannot be expressed in an explicit regression function form; Random forest gives an averaged prediction and fails to recognize the multiple patterns at one position. For Example 4, only the resolution-wise regression successfully distinguishes the two subgroups of the heterogeneous mean with homogeneous error model and the homogeneous mean with a heteroscedastic error model.

Based on the simulation results, we can see that resolution-wise regression has a better distribution prediction when the variance σ^2 of the error is larger, which seems paradoxical with the point prediction of some common regression methods. In fact, for point prediction, the loss comes from the random error term. A large variance leads to a large loss. However, for distribution prediction, the loss comes from the accumulated probability of possible response values.

The variance affects the shape of the distribution of the response. For a smaller variance, the data are concentrated, and more precise resolutions are needed, which bring the difficulties for the estimation of the corresponding logistic regressions. Hence, for the same expansion order d_Y, our method gives less accurate results on smaller variances. This also shows some insights for a general prediction problem that distribution prediction is a good alternative to capture the whole picture if the random error is relatively large.

6. Real Data Analysis

We analyze the real estate valuation dataset (Yeh and Hsu 2018), obtained from UCI machine learning Repository (*https://archive.ics.uci.edu*). The dataset contains the unit-area price and the corresponding six explanatory variables of 414 houses collected from Sindian District, New Taipei City. The six explanatory variables include the transaction date, the house age, the distance to the nearest MRT station, the number of convenience stores in the living circle on foot, the latitude, and the longitude. We are interested in how the house price can be explained by these variables. We consider three methods: (a) Naive method with the small neighborhood of the nearest 10 samples, where the distance is measured by the Mahalanobis

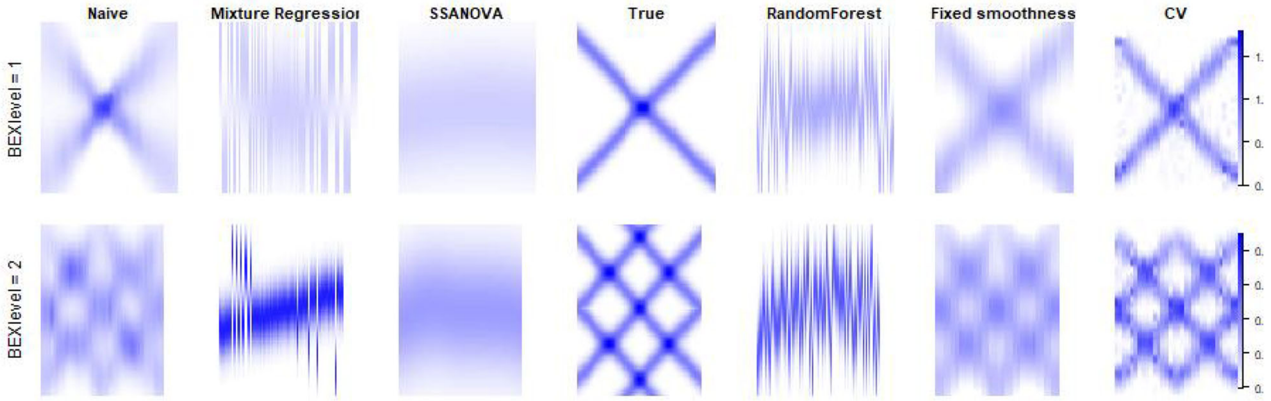

Figure 6. Heatmaps of prediction distributions for Example 1 with respect to one or multiple crosses and six methods: naive method, regression mixture, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV, from left to right, respectively. A darker color indicates a larger PDF value at the corresponding predicted response.

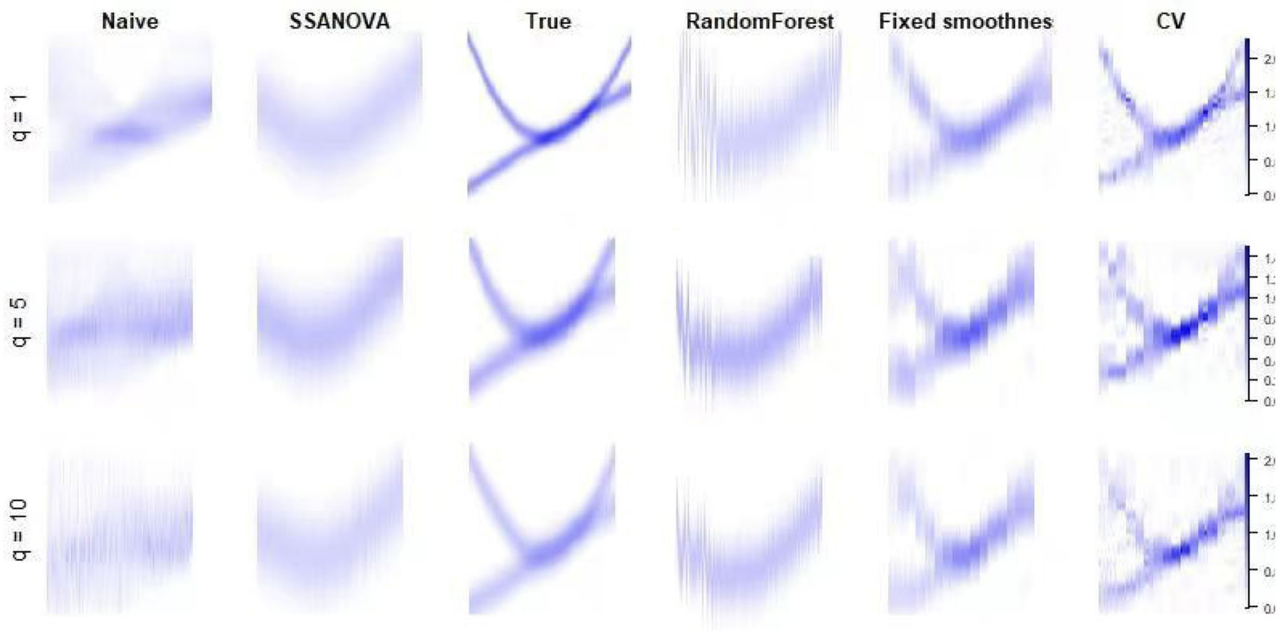

Figure 7. Heatmaps of prediction distributions for Example 2 with respect to dimension $q = 1, 5, 10$ and five methods: naive method, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV, from left to right, respectively. A darker color indicates a larger PDF value at the corresponding predicted response.

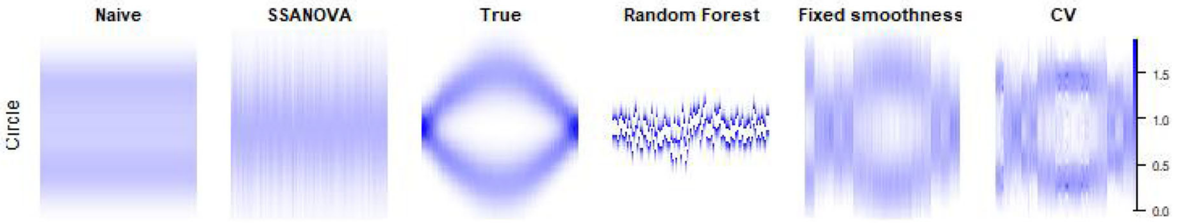

Figure 8. Heatmaps of prediction distributions for Example 3 with respect to circular implicit functional relationship and five methods: naive method, SSANOVA, Random Forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV, from left to right, respectively. A darker color indicates a larger PDF value at the corresponding predicted response.

distance (Rosenbaum 1995) of the rank vector within each predictor; (b) SSANOVA with cubic splines on each main effects and interactions; (c) Resolution-wise regression with $d_X = 5$ and $d_Y = 5$.

There are three interesting results we find from the application of resolution-wise regression.

House prices rely on some features through a nonlinear relationship. Such a nonlinear pattern can be detected by the

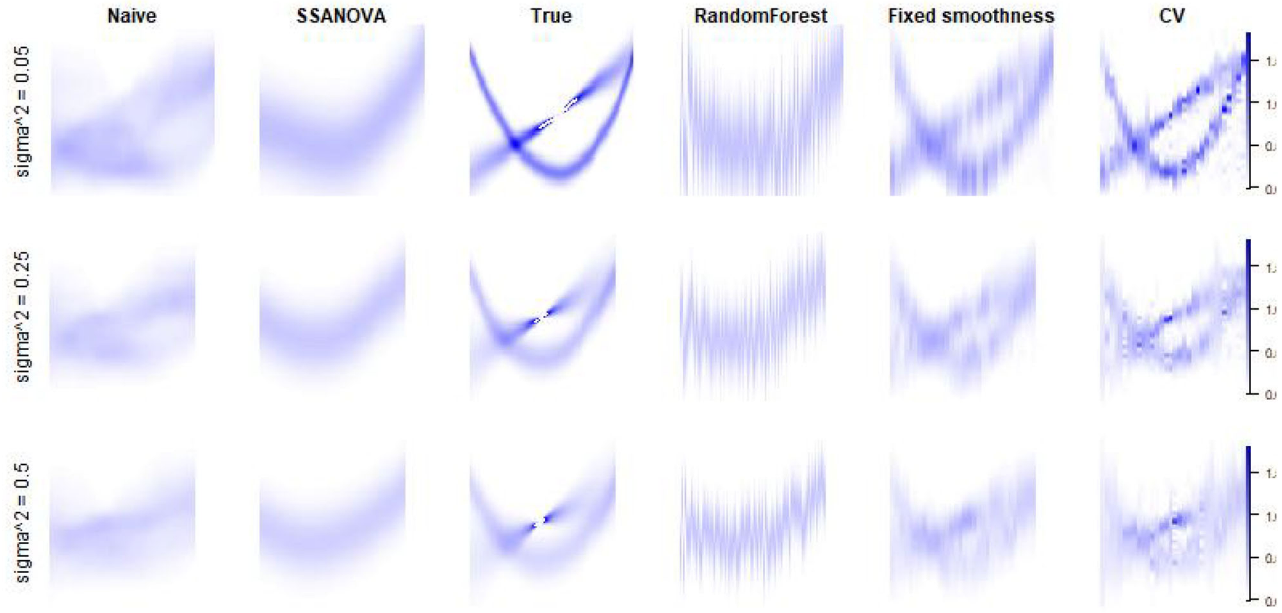

Figure 9. Heatmaps of prediction distributions for Example 4 with respect to different variances of random errors and five methods: naive method, SSANOVA, random forest, resolution-wise regression with fixed smoothness, and resolution-wise regression with CV, from left to right, respectively. A darker color indicates a larger PDF value at the corresponding predicted response.

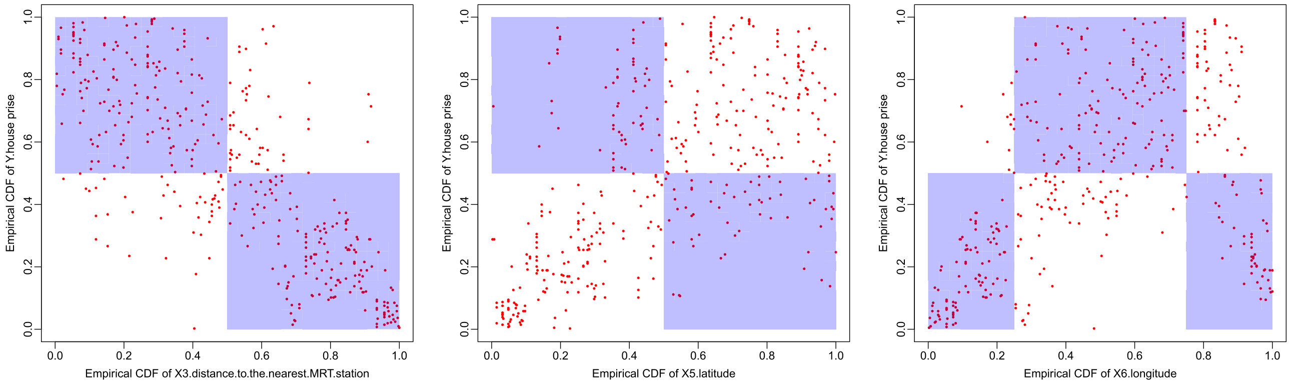

Figure 10. Relevant variables and the corresponding most significant patterns. For the distance to the nearest MRT station and the latitude, there exists a linear relationship to the housing prices. For the longitude, the most asymmetric interaction is $A_1A_2B_1$, which implies a nonlinear dependence.

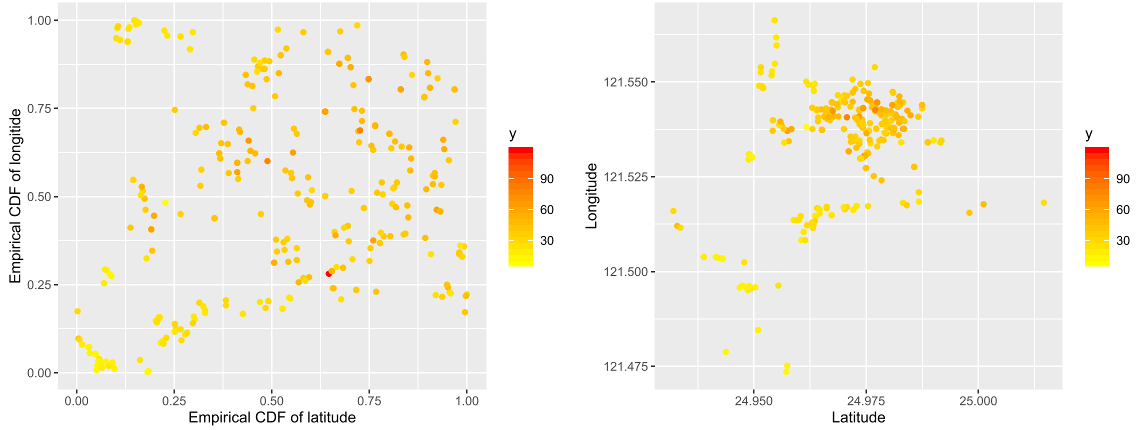

Figure 11. Interaction of latitude and longitude. The high-price houses concentrate around the downtown area.

BET screening, as shown in Figures 10 and 11. The latitude and the longitude show skewed quadratic effects which are captured by the relevant patterns in depth 1 and depth 2, respectively. For the screening of interaction patterns, all the C_6^2 interactions are significant. As an example shown in Figure 11, the interaction of the latitude and the longitude shows a concentration of high-price houses, which can be pinpointed around the downtown area. There is also some linear pattern found in the data. As shown in Figure 10, the distance to the nearest MRT station and the number of convenience stores in the living circle on foot has a linear effect on the house price which is captured by the relevant pattern in depth 1.

Potential heterogeneity can be detected by resolution-wise regression, as shown in Figure 12. The proposed method clearly predicts the house price concentrating on two groups around 30 and 60, which is new information that is not provided by existing methods. For examples, the SSANOVA completely misses such heterogeneity. The naive method provides a distribution which vaguely suggests probability mass toward the right tail. However, the subgroup information identified by the naive method is not very clear (Figure 12).

The detected heterogeneity can also be demonstrated through the additional information from the map of the city. As shown in Figure 13, the nearest 10 samples measured by the Mahalanobis distance form two groups classified by the river and the highway. The two groups differ in the house price, and both contribute to the distribution estimation. The effect of the river and the highway play the role of unobserved variable, which leads to the bimodal shape of the prediction distribution. The prediction from our method suggests that the price of the particular house is more likely to be close to the three houses on the lower-left side of the river, which has an average price of 25.75. This is verified as correct prediction through the actual map. In particular, the location of the house is indeed on the lower-left side of the river. This example thus illustrates the advantage of the proposed method. Specifically, it can detect heterogeneity in the data and provide accurate probability statements about subgroup information.

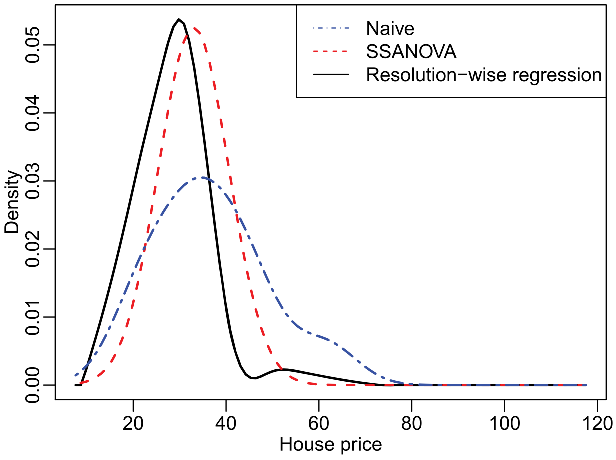

Figure 12. Predicted distributions by naive method by the nearest 10 samples, SSANOVA, and resolution-wise regression.

In summary, this real data analysis indicates that resolution-wise regression model can capture the heterogeneous pattern, thus, can deliver more detailed prediction information than traditional methods. Since no distribution assumption is required, our method is rather general and robust.

7. Conclusion

In this article, we propose resolution-wise regression model to predict the distribution of the response with heterogeneous data. The complicated relationship between the response and the explanatory variables can be decomposed into the relationship of resolutions of the response and patterns of the predictors based on binary expansions. A set of penalized logistic regressions establish the effect of patterns having on the resolutions. By BID transformation, our method can estimate the cell probability of the histogram of the response, which is an approximation of the distribution of the response. We also show the consistency

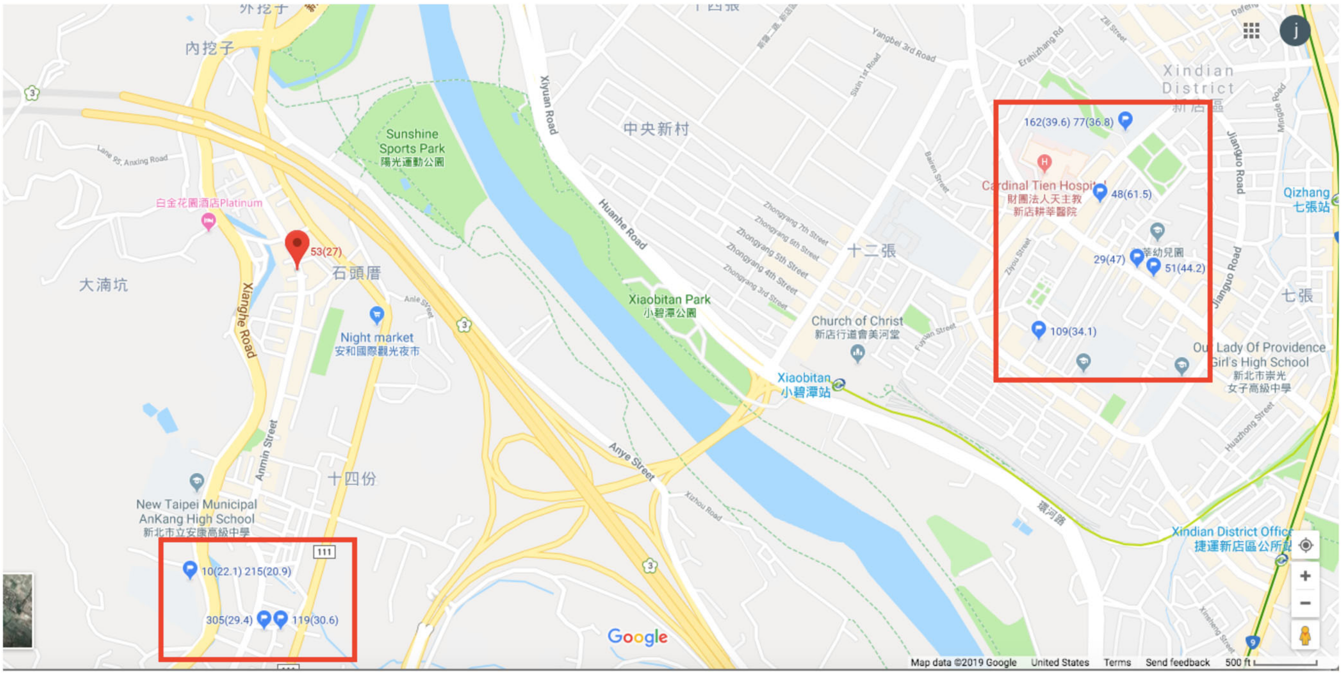

Figure 13. The testing sample (red pin) and the nearest 10 samples measured by the Mahalanobis distance of the rank vector within every predictor (blue pins, where two samples on the power left side have the same location, and two samples on the lower left side have the same location) on the map. In these 10 houses, four are on the lower left side of the river with an average price of 25.75, and six are on the upper right side of river with an average price of 43.87.

of the cell probabilities. Numerical studies demonstrate the effectiveness of the proposed method.

Appendix

Proof of Theorem 1:

We split the whole proof into two steps: (a) the selected variables from the BET screening are equivalent to those from the sure independence screening based on the MMLE, (b) the proposed logistic regression satisfies the conditions in Fan and Song (2010) to achieve the sure independence screening property.

(1). The BET test statistic

$$\begin{aligned} S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} &= \left| \sum_{i=1}^{n} I(\dot{A}_{(j),i}\dot{B}_{(m),i} = 1) - \sum_{i=1}^{n} I(\dot{A}_{(j),i}\dot{B}_{(m),i} = -1) \right| \\ &= \left| \sum_{i=1}^{n} \dot{A}_{(j),i}\dot{B}_{(m),i} \right| \\ &= \sqrt{n} \left| \sum_{i=1}^{n} \tilde{A}_{(j),i}\dot{B}_{(m),i} \right|. \end{aligned}$$

By the definition in (12), and $\tilde{B}_{(m)} = (\dot{B}_{(m)} + 1)/2$, the MMLE can be obtained by the optimization with respect to $\dot{B}_{(m),i}$'s, that is,

$$\begin{aligned} \tilde{\beta}_{m,j} = \arg\min_{\beta_{m,j}} \frac{1}{n} \sum_{i=1}^{n} -\frac{\dot{B}_{(m),i} + 1}{2} \tilde{A}_{(j),i}\beta_{m,j} \\ + \log(1 + e^{\tilde{A}_{(j),i}\beta_{m,j}}),\ j = 1, \ldots, L. \end{aligned}$$

Setting the derivative of the above objective function with respect to $\beta_{m,j}$ to be zero, we have that $\tilde{\beta}_{m,j}$ satisfies

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \frac{\dot{B}_{(m),i}\tilde{A}_{(j),i}}{2} &= \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\tilde{A}_{(j),i} e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}{1 + e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}} - \frac{\tilde{A}_{(j),i}}{2} \right) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{A}_{(j),i} e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}{1 + e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}. \end{aligned}$$

The second equation holds since there are $n/2$ samples with $\tilde{A}_{(j),i} = 1/\sqrt{n}$ and the other $n/2$ samples with $\tilde{A}_{(j),i} = -1/\sqrt{n}$, due to the binary expansion from the empirical CDF transformation. Denote $t(\tilde{\beta}_{m,j}) := \sum_{i=1}^{n} \frac{\tilde{A}_{(j),i} e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}{1+e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}$. Differentiating with respect to $\tilde{\beta}_{m,j}$, we obtain $\frac{dt(\tilde{\beta}_{m,j})}{d\tilde{\beta}_{m,j}} = \sum_{i=1}^{n} \frac{\tilde{A}^2_{(j),i} e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}}}{(1+e^{\tilde{A}_{(j),i}\tilde{\beta}_{m,j}})^2} > 0$. Hence, $t(\tilde{\beta}_{m,j})$ is strictly increasing with respect to $\tilde{\beta}_{m,j}$. For $\gamma_{m,n} > 0$, if $\tilde{\beta}_{m,j} \geqslant \gamma_{n,m}$, $t(\tilde{\beta}_{m,j}) \geqslant t(\gamma_{n,m})$. If $\tilde{\beta}_{m,j} \leqslant -\gamma_{n,m}$,

$$\begin{aligned} t(\tilde{\beta}_{m,j}) \leqslant t(-\gamma_{n,m}) &= \sum_{i=1}^{n} \frac{\tilde{A}_{(j),i} e^{-\tilde{A}_{(j),i}\gamma_{n,m}}}{1 + e^{-\tilde{A}_{(j),i}\gamma_{n,m}}} = \sum_{i=1}^{n} \frac{\tilde{A}_{(j),i}}{1 + e^{\tilde{A}_{(j),i}\gamma_{n,m}}} \\ &= \sum_{i=1}^{n} \left(\tilde{A}_{(j),i} - \frac{\tilde{A}_{(j),i} e^{\tilde{A}_{(j),i}\gamma_{n,m}}}{1 + e^{\tilde{A}_{(j),i}\gamma_{n,m}}} \right) \\ &= -t(\gamma_{n,m}). \end{aligned}$$

Since $S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} = 2\sqrt{n}|t(\tilde{\beta}_{m,j})|$, we have $S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} \geqslant 2\sqrt{n}\gamma_{n,m}$. Hence, we have the variable selection index set$\{1 \leqslant j \leqslant L : S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} \geqslant 2\sqrt{n}t(\gamma_{n,m})\} \supseteq \{1 \leqslant j \leqslant L : |\tilde{\beta}_{m,j}| \geqslant \gamma_{n,m}\}$. Similarly, we have$\{1 \leqslant j \leqslant L : S_{\boldsymbol{a}_{(j)}\boldsymbol{b}_{(m)}} \geqslant 2\sqrt{n}t(\gamma_{n,m})\} \subseteq \{1 \leqslant j \leqslant L : |\tilde{\beta}_{m,j}| \geqslant \gamma_{n,m}\}$. Denote $I_j^+ := \{1 \leqslant i \leqslant n : \tilde{A}_{(j),i} = 1/\sqrt{n}\}$, $I_j^- := \{1 \leqslant i \leqslant n : \tilde{A}_{(j),i} = -1/\sqrt{n}\}$. Taking $\gamma_{n,m} = c_{4,m}n^{-\kappa_m}$ for some $c_{4,m} > 0$, we have

$$\begin{aligned} 2\sqrt{n}t(\gamma_{n,m}) &= 2\sqrt{n} \sum_{i=1}^{n} \frac{\tilde{A}_{(j),i} e^{\tilde{A}_{(j),i} c_{4,m} n^{-\kappa_m}}}{1 + e^{\tilde{A}_{(j),i} c_{4,m} n^{-\kappa_m}}} \\ &= 2\sqrt{n} \sum_{i \in I_j^+} \frac{\frac{1}{\sqrt{n}} e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}}{1 + e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} \\ &\quad - 2\sqrt{n} \sum_{i \in I_j^-} \frac{\frac{1}{\sqrt{n}} e^{-c_{4,m} n^{-\frac{1}{2}-\kappa_m}}}{1 + e^{-c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} \\ &= \frac{n e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}}{1 + e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} - \frac{n}{1 + e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} \\ &= \frac{n(e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}} - 1)}{1 + e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} = \frac{nO(n^{-\frac{1}{2}-\kappa_m})}{1 + e^{c_{4,m} n^{-\frac{1}{2}-\kappa_m}}} \\ &= O(n^{\frac{1}{2}-\kappa_m}). \end{aligned}$$

Hence, the BET screening is equivalent to the sure independence screening based on MMLE. This ensures that the estimation methods are the same as those in Fan and Song (2010).

(2). Under the binary expansion, the variables are bounded in [0, 1] after empirical CDF transformation, so that the conditions $A-C$ in Fan and Song (2010) are naturally satisfied. Assumption 1 is analogous to Condition *E*. So we only need to check Condition *D*. For the proposed logistic regression, let $w_0 = 1$, and we have

$$\begin{aligned} &\mathbb{E}\exp(\log(1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0 + w_0}) - \log(1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0})) \\ &\quad + \mathbb{E}\exp(\log(1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0 - w_0}) - \log(1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0})) \\ &= \mathbb{E}\frac{1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0 + w_0}}{1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0}} + \mathbb{E}\frac{1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0 - w_0}}{1 + e^{\tilde{\boldsymbol{A}}^T\beta_m^0}} \\ &= 2 + \mathbb{E}\frac{(e^{w_0/2} - e^{-w_0/2})^2}{1 + e^{-\tilde{\boldsymbol{A}}^T\beta_m^0}} \\ &\leqslant 2 + (e^{w_0/2} - e^{-w_0/2})^2 = 2 + (e^{1/2} - e^{-1/2})^2. \end{aligned}$$

Taking $w_1 = 2, h_{1,m} = 3, h_{0,m} = 1, \alpha = 1$ satisfies Condition *D*.

Hence, by Theorem 4 in Fan and Song (2010), for $\delta_{n,m} = O(n^{\frac{1}{2}-\kappa_m})$, the sure independence screening property is achieved.

Proof of Theorem 2: For logistic regression, the function $b(\cdot)$ of Condition G in Fan and Song (2010) is $b(x) = \log(1 + e^x)$. Since $0 < \frac{d^2b(x)}{dx^2} = \frac{e^{-x}}{(1+e^{-x})^2} < 1$, Condition G in Fan and Song (2010) is met. Together with Assumption 2, by Theorem 5 in Fan and Song (2010), the proof is completed.

Let $\mathcal{F}$ be a normed real vector space. For the *m*th logistic regression function $f_m \in \mathcal{F} : \{-1, 1\}^r \to \mathbb{R}$, where $r := \max_{1 \leqslant m \leqslant 2^{d_Y}-1} |\mathcal{M}_{m,\delta_{n,m}}|$, the negative log-likelihood loss $\rho_{f_m} : \{-1, 1\}^{r+1} \to \mathbb{R}$ is $\rho_{f_m}(Z_m) = \rho_{f_m}(\dot{\boldsymbol{A}}^S, \dot{B}_{(m)}) = \log(1 + e^{-\dot{B}_{(m)}f(\dot{\boldsymbol{A}}^S)})$, $m = 1, \ldots, 2^{d_Y}-1$, where $\dot{\boldsymbol{A}}^S := (\dot{A}^S_{(1)}, \ldots, \dot{A}^S_{(r)})^T$ is the predictor vector including the selected r patterns. For a loss function ρ_{f_m}, define the empirical risk for ρ_{f_m} by $P_n\rho_{f_m} := \frac{1}{n}\sum_{i=1}^{n} \rho_{f_m}(Z_{m,i})$, and the theoretical risk by $P\rho_{f_m} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\rho_{f_m}(Z_{m,i})$. Consider the collection $\mathcal{F}$ to be a linear-model class, that is, $\mathcal{F} := \{f^\beta : \beta \in \mathbb{R}^p\}$, where $\beta \mapsto f_\beta$ is linear. For the mth regression, note that the true coefficient vector β_m^0 is the minimizer of the theoretical risk

$$\beta_m^0 := \arg\min_{\beta} P\rho_{f_m^\beta}, \tag{13}$$

and $f_m^0 := f_m^{\beta_m^0}$. We assume for simplicity that the minimum exists and unique. For $f_m^\beta \in \mathcal{F}$, the excess risk is defined by $\mathcal{E}(f_m^\beta) := P(\rho_{f_m^\beta} - \rho_{f_m^0})$. The lasso estimator is $\hat{\beta}_m = \arg\min_\beta\{P_n\rho_{f_m^\beta} + \lambda_m||\beta||_1\}$, $m = 1, \ldots, 2^{d_Y} - 1$, where $||\cdot||_1$ is the ℓ_1-norm and λ_m is a tuning parameter. The estimation of the regression function is $\hat{f}_m = f_m^{\hat{\beta}_m}$.

Denote $\pi_m(\dot{\boldsymbol{A}}^S) = P(\dot{B}_{(m)} = 1|\dot{\boldsymbol{A}}^S)$ and $e_m(\dot{\boldsymbol{A}}^S) = \mathbb{E}(\dot{B}_{(m)}|\dot{\boldsymbol{A}}^S)$. We have $e_m(\dot{\boldsymbol{A}}^S) = 2\pi_m(\dot{\boldsymbol{A}}^S) - 1$. Hence, based on the link function of logistic regression, we can define a functional g mapping $e_m(\cdot)$ to the regression function $f_m^\beta(\cdot)$:

$$g(e_m)(\cdot) := f_m^\beta(\cdot) = \log\left(\frac{\pi_m(\cdot)}{1 - \pi_m(\cdot)}\right) = \log\left(\frac{\frac{e_m(\cdot)+1}{2}}{1 - \frac{e_m(\cdot)+1}{2}}\right). \tag{14}$$

Denoting $e_m^0(\dot{\boldsymbol{A}}^S)$ as the true expectation corresponding to β_m^0 , by (14), we have $f_m^0 = g(e_m^0)$. Similarly, recall that $\hat{e}_m(\dot{\boldsymbol{A}}^S)$ is the estimated expectation, and thus we have $\hat{f}_m = g(\hat{e}_m)$.

For a given index set $S_m \subset \{1, \ldots, r\}$, define $\beta_{m,j,S_m} := \beta_{m,j}1\{j \in S_m\}$, $m = 1, \ldots, 2^{d_Y} - 1$, $j = 1, \ldots, r$. Denote the estimator restricted to $\beta_{m,S_m} = (\beta_{m,1,S_m}, \ldots, \beta_{m,r,S_m})^T$ by $\hat{\beta}_{m,S_m} := \arg\min_{\beta=\beta_{m,S_m}} P_n\rho_{f_m^\beta}$. Write $\hat{f}_{m,S} := f_m^{\hat{\beta}_{m,S_m}}$. Restricted to β_{m,S_m}'s, the best approximation of f_m^0 is $f_{m,S}^0 := f_m^{\beta_{m,S_m}^0}$, where $\beta_{m,S_m}^0 := \arg\min_{\beta=\beta_{m,S_m}} P\rho_{f_m^\beta}$.

The following assumption requires a certain compatibility of ℓ_1-norm with the norm on $\mathcal{F}$, which is a regular assumption for the theoretical framework for lasso.

Assumption 3. (Compatibility condition) We say that the compatibility condition is met for the set S_m with constant $\phi_m > 0$, if for all β_m satisfying $||\beta_{m,S_m^c}||_1 \leqslant 3||\beta_{m,S_m}||_1$, it holds that $||\beta_{m,S_m}||_1^2 \leqslant ||f_m||^2 s_m/\phi_m^2$.

Next, we show the definition of the margin condition (Bülmann and van de Geer 2011) and demonstrate that the penalized logistic regression satisfies the condition with a quadratic margin.

Definition 1 (Margin condition). Denote a "neighborhood" of $f_m^0 \in \mathcal{F}$ by $\mathcal{F}_{\eta_m} := \{f \in \mathcal{F} : ||f - f_m^0||_\infty \leqslant \eta_m\}$ with constant $\eta_m > 0$. We say that the margin condition holds with a strictly convex function G, if for all $f \in \mathcal{F}_{\eta_m}$, we have $\mathcal{E}(f) \geqslant G(||f - f_m^0||)$, where $||\cdot||$ is the norm defined on $\mathcal{F}$.

Assumption 4. For any fixed $\dot{\boldsymbol{A}}^S$, there exists some constant $0 \leqslant \varepsilon_m^0 \leqslant 1$ such that $\varepsilon_m^0 \leqslant \pi_m(\dot{\boldsymbol{A}}^S) \leqslant 1 - \varepsilon_m^0$, $m = 1, \ldots, 2^{d_Y} - 1$.

Lemma 5. Under Assumption 4, the margin condition holds for all $2^{d_Y} - 1$ penalized logistic regressions with a quadratic margin, that is, $G_m(u) = c_m u^2$ for the mth regression.

The technical proof of Lemma 5 can be found in supplementary materials. For the mth regression, the oracle β_m^* (Bülmann and van de Geer 2011) is defined by

$$\beta_m^* := \arg\min_{\beta:S_\beta\in\Psi}\left\{3\mathcal{E}(f_m^\beta) + \frac{8\lambda_m^2 s_\beta}{c_m\phi_m^2}\right\}, \tag{15}$$

where $S_\beta := \{j : \beta_j \neq 0\}$, $s_\beta := |S_\beta|$ denotes the cardinality of S_β, ϕ_m^2 is a compatibility constant, and Ψ is a suitable large collection of index sets. Denote the index set of nonzero coefficients by $S_m^0 := \{j : \beta_{m,j}^0 \neq 0\}$, and the cardinality of S_m^0 by $s_m := |S_m^0|$. Assuming f_m^0 is linear, we can take $\Psi = \{S_m^0\}$. Hence, the definition of β_m^* is consistent with the definition of β_m^0 in (13), since the second term of (15) does not rely on β. In this context, we only use the notation β_m^0. Denote the minimum of (15) by $2\epsilon_m^* := 3\mathcal{E}(f_m^0) + \frac{8\lambda_m^2 s_m}{c_m\phi_m^2}$. Define $Z_{M_m} := \sup_{||\beta-\beta_m^0||\leqslant M_m}|\nu_n(\beta) - \nu_n(\beta_m^0)|$, where $\nu_n(\beta_m) := (P_n - P)\rho_{f_m^{\beta_m}}$ is the empirical process. Set $M_m^* := \epsilon_m^*/\lambda_m^0$, and $\mathcal{T}_m := \{Z_{M_m^*} \leqslant \lambda_m^0 M_m^*\} = \{Z_{M_m^*} \leqslant \epsilon_m^*\}$. Bülmann and van de Geer (2011) showed that one can choose $\lambda_m^0 \asymp \sqrt{\log(n^{2\kappa_m})/n}$ such that the set $\mathcal{T}_m$ holds with large probability.

Assumption 5. For some constant $\eta_m > 0$, $f_m^{\beta_m} \in \mathcal{F}_{\eta_m} := \{||f_m^{\beta_m} - f_m^0||_\infty \leqslant \eta_m\}$ for all $||\beta_m - \beta_m^0||_1 \leqslant M^*$, as well as $f_m^0 \in \mathcal{F}_{\eta_m}$.

According to the BID equation, we estimate $\boldsymbol{p}$ by solving the optimization (9). From the optimization, $\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}}$ is an approximation of $\hat{e}_m$, where $\boldsymbol{H}_{m+1}$ is the $(m+1)$th row of $\boldsymbol{H}$, since $\hat{e}_m$ is the $(m+1)$th entry of $\hat{\boldsymbol{E}}$. Hence, $g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})$ is the estimated m-th regression function corresponding to $\hat{\boldsymbol{p}}$. The following theorem gives the consistency of cell probability vector $\hat{\boldsymbol{p}}$ in terms of excess risk of $g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})$.

We now turn to the proof of Theorem 3. Below is its statement again for convenience.

Theorem 4. Assume Assumptions 1–5 hold, where Assumption 3 holds with the set S_m^0. For the logistic regression with covariates corresponding to the BET screening set $\mathcal{M}_{m,\delta_{n,m}}$, suppose that λ_m satisfies $\lambda_m \geqslant 8\lambda_m^0$. Then on the set $\mathcal{T}_m$, we have,

$$\begin{aligned}&\mathcal{E}(g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})) + \lambda_m||\hat{\beta}_m - \beta_m^0||_1\\ &\quad\leqslant 6\mathcal{E}(f_m^0) + \frac{16\lambda_m^2 s_m}{c_m\phi_m^2} + \frac{32\lambda K s 2^{d_Y}}{c\phi^2},\end{aligned}$$

where $K > 0$ is a constant, $\lambda = \max_{1\leqslant m\leqslant 2^{d_Y}-1}\lambda_m$, $s = \max_{1\leqslant m\leqslant 2^{d_Y}-1} s_m$, $c = \min_{1\leqslant m\leqslant 2^{d_Y}-1} c_m$, $\phi = \min_{1\leqslant m\leqslant 2^{d_Y}-1} \phi_m$.

Before the proof of Theorem 3, we first state the oracle inequality for penalized logistic regression by Bülmann and van de Geer (2011) as follows.

Lemma 6. Assume Assumptions 3–5 hold, where 3 holds with the set S_m^0. Suppose that λ_m satisfies the inequality $\lambda_m \geqslant 8\lambda_m^0$. Then on the set $\mathcal{T}_m$, we have

$$\mathcal{E}(\hat{f}_m) + \lambda_m||\hat{\beta}_m - \beta_m^0||_1 \leqslant 6\mathcal{E}(f_m^0) + \frac{16\lambda_m^2 s_m}{c_m\phi_m^2},$$

where $c_m = (\frac{e^{\eta_m}}{\varepsilon_m^0} + 1)^{-2}$.

Then we have the proof for Theorem 3 as follows.

Proof of Theorem 3: For the excess risk of $g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})$, we have

$$\begin{aligned}\mathcal{E}(g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})) &= P\rho_{g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})} - P\rho_{f_m^0} = P\rho_{g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})} - P\rho_{g(e_m^0)}\\ &= (P\rho_{g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})} - P\rho_{g(\hat{e}_m)}) + (P\rho_{g(\hat{e}_m)} - P\rho_{g(e_m^0)})\\ &= (P\rho_{g(\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}})} - P\rho_{g(\hat{e}_m)}) + \mathcal{E}(\hat{f}_m) \overset{\triangle}{=} I + II.\end{aligned}$$

It can be shown that the function $P\rho_{g(e_m)}$ is Lipschitz continuous, with the Lipschitz constant K_m obtained from the first derivative

$$\left|\frac{dP\rho_{g(e_m)}}{de_m}\right| = \left|\left(\frac{e^{f_m}}{1+e^{f_m}} - \pi_m\right)\left(\frac{1}{1+e_m} + \frac{1}{1-e_m}\right)\right| \leqslant \frac{2(1-\varepsilon_m^0)}{1-1(1-2\varepsilon_m^0)^2} \triangleq K_m.$$

For part I, denoting the true expectation vector by $\boldsymbol{E}^0$ and the corresponding true cell probability vector by $\boldsymbol{p}^0 = \boldsymbol{H}^{-1}\boldsymbol{E}^0$, we have

$$|I| \leqslant K_m|\boldsymbol{H}_{m+1}\hat{\boldsymbol{p}} - \hat{e}_m| \leqslant K_m||\boldsymbol{H}\hat{\boldsymbol{p}} - \hat{\boldsymbol{E}}||_1 \leqslant K_m||\boldsymbol{H}\boldsymbol{p}^0 - \hat{\boldsymbol{E}}||_1 \leqslant K_m(||\boldsymbol{H}\boldsymbol{p}^0 - \boldsymbol{E}^0||_1 + ||\boldsymbol{E}^0 - \hat{\boldsymbol{E}}||_1).$$

By Lemma 6, $||\hat{\beta}_m - \beta_m^0||_1 \leqslant \frac{16\lambda_m s_m}{c_m\phi_m^2}$. Since $|\hat{f}_m - f_m^0| \leqslant ||\hat{\beta}_m - \beta_m^0||_1$ with predictors taking values from $\{-1, 1\}$, we have $|\hat{f}_m - f_m^0| \leqslant \frac{16\lambda_m s_m}{c_m\phi_m^2}$. One can similarly show that the function $g^{-1}(\cdot)$ is Lipschitz continuous with Lipschitz constant two. Hence, we have $|\hat{e}_m - e_m^0| \leqslant 2|\hat{f}_m - f_m^0| \leqslant \frac{32\lambda_m s_m}{c_m\phi_m^2}$, and thus $||\boldsymbol{E}^0 - \hat{\boldsymbol{E}}||_1 \leqslant \frac{32\lambda s 2^{d_Y}}{c\phi^2}$, where $\lambda = \max_{1\leqslant m\leqslant 2^{d_Y}-1} \lambda_m$, $s = \max_{1\leqslant m\leqslant 2^{d_Y}-1} s_m$, $c = \min_{1\leqslant m\leqslant 2^{d_Y}-1} c_m$, and $\phi = \min_{1\leqslant m\leqslant 2^{d_Y}-1} \phi_m$.

For the true expectation vector and the true cell probability vector, we have $||\boldsymbol{H}\boldsymbol{p}^0 - \boldsymbol{E}^0||_1 = 0$. Denote $K = \max_{1\leqslant m\leqslant 2^{d_Y}-1} K_m$. Together with the inequality in Lemma 6 to handle the part II, the proof is completed.

Supplementary Materials

Supplement materials of this article include additional simulation studies, additional proofs and R code.

Acknowledgments

The authors would like to thank the editor, the associate editor, and two anonymous referees for their valuable comments and suggestions.

Disclosure Statement

The authors declare that they have no competing financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding

Research of Qizhai Li was supported in part by CAS Project for Young Scientists in Basic Research (Grant No. YSBR-034). Kai Zhang was supported in part by NSF grants DMS-1613112, IIS-1633212, DMS-1916237, and DMS-2152289. Yufeng Liu was supported in part by NSF grant DMS-2100729 and NIH grant R01GM126550.

ORCID

Kai Zhang http://orcid.org/0000-0002-4791-880X
Yufeng Liu http://orcid.org/0000-0002-1686-0545

References

Bülmann, P., and van de Geer, S. (2011), *Statistics for High-Dimensional Data*, Berlin: Springer. [15]

Chen, J., Tran-Dinh, Q., Kosorok , M. R., and Liu, Y. (2021), "Identifying Heterogeneous Effect using Latent Supervised Clustering with Adaptive Fusion," *Journal of Computational and Graphical Statistics*, 30, 43–54. [2]

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with NP-dimensionality," *Annals of Statistics*, 38, 3567–3604. [2,7,8,14]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [2]

Green, P., and Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, London: Chapman and Hall. [1]

Greenshtein, E., and Ritov, Y. (2004), "Persistence in High-Dimensional Linear Predictor Selection and the Virtue of Overparametrization," *Bernoulli*, 10, 971–988. [8]

Gu, C. (2002), *Smoothing Spline ANOVA Models*, New York: Springer. [1]

Guo, F. J., Levina, E., Michailidis, G., and Zhu, J. (2010), "Pairwise Variable Selection for High-Dimensional Model-Based Clustering," *Biometrics*, 66, 793–804. [2]

Hocking, T., Joulin, A., Bach, F., and Vert, J. P. (2011), "Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties," in *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*, eds. L. Getoor and T. Scheffer, pp. 745–52, New York: Omnipress. [2]

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991), "Adaptive Mixtures of Local Experts," *Neural Computation*, 3, 79–87. [2]

Kac, M. (1959), *Statistical Independence in Probability, Analysis and Number Theory*, Washington DC: Mathematical Association of America. [3]

Lindsten, F., Ohlsson, H., and Ljung, L. (2011), "Clustering using Sum-of-Norms Regularization: With Application to Particle Filter Output Computation," in *2011 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 201–204. [2]

Ma, S., and Huang, J. (2017), "A Concave Pairwise Fusion Approach to Subgroup Analysis," *Journal of the American Statistical Association*, 112, 410–423. [2]

Pan, W., and Shen, X. (2006), "Penalized Model-based Clustering with Application to Variable Selection," *Journal of Machine Learning Research*, 8, 1145–1164. [2]

Pan, W., Shen, X., and Liu, B. (2013), "Cluster Analysis: Unsupervised Learning via Supervised Learning with a Non-convex Penalty," *Journal of Machine Learning Research*, 14, 1865–1889. [2]

Raftery,A., and Dean, N. (2006), "Variable Selection for Model-based Clustering," *Journal of the American Statistical Association*, 101, 168–178. [2]

Rosenbaum, P. (1995), *Observational Studies*, New York: Springer. [11]

Sylvester, J. J. (1867), "LX. Thoughts on Inverse Orthogonal Matrices, Simultaneous Signsuccessions, and Tessellated Pavements in Two or More Colours, with Applications to Newton's Rule, Ornamental Tile-Work, and the Theory of Numbers," *The London, Edinburgh, and Dublin Philosophical Magazine*, 34, 461–475. [4]

Tang, X., and Qu, A. (2017), "Individualized Multi-Directional Variable Selection," arXiv: 1709.05062. [2]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [4]

Wahba, G. (1990), *Spline Models for Observational Data*, Philadelphia: SIAM. [1]

Yeh, I. C., and Hsu, T. K. (2018), "Building Real Estate Valuation Models with Comparative Approach through Case-based Reasoning," *Applied Soft Computing*, 65, 260–271. [10]

Zhang, K. (2019), "BET on Independence," *Journal of the American Statistical Association*, 114, 1620–1637. DOI: 10.1080/01621459.2018.1537921. [2,4,5]

Zhang, K., Zhao, Z., and Zhou, W. (2021), "BEAUTY Powered BEAST," arXiv: 2103.00674. [9]