Generalized Hierarchical Sparse Model for Arbitrary-Order Interactive Antigenic Sites Identification in Flu Virus Data

Lei Han[†], Yu Zhang[‡], Xiu-Feng Wan[§], Tong Zhang[†¶]
[†]Department of Statistics, Rutgers University
[‡]Department of Computer Science and Engineering, Hong Kong University of Science and Technology
[§]College of Veterinary Medicine, Mississippi State University
[¶]Baidu Inc. Beijing, China
lhan@stat.rutgers.edu; zhangyu@cse.ust.hk; wan@cvm.msstate.edu; tzhang@stat.rutgers.edu

ABSTRACT

Recent statistical evidence has shown that a regression model by incorporating the interactions among the original covariates (features) can significantly improve the interpretability for biological data. One major challenge is the exponentially expanded feature space when adding high-order feature interactions to the model. To tackle the huge dimensionality, Hierarchical Sparse Models (HSM) are developed by enforcing sparsity under heredity structures in the interactions among the covariates. However, existing methods only consider pairwise interactions, making the discovery of important high-order interactions a non-trivial open problem. In this paper, we propose a Generalized Hierarchical Sparse Model (GHSM) as a generalization of the HSM models to learn arbitrary-order interactions. The GHSM applies the ℓ_1 penalty to all the model coefficients under a constraint that given any covariate, if none of its associated kth-order interactions contribute to the regression model, then neither do its associated higher-order interactions. The resulting objective function is non-convex with a challenge lying in the coupled variables appearing in the arbitrary-order hierarchical constraints and we devise an efficient optimization algorithm to directly solve it. Specifically, we decouple the variables in the constraints via both the GIST and ADMM methods into three subproblems, each of which is proved to admit an efficiently analytical solution. We evaluate the GHSM method in both synthetic problem and the antigenic sites identification problem for the flu virus data, where we expand the feature space up to the 5th-order interactions. Empirical results demonstrate the effectiveness and efficiency of the proposed method and the learned high-order interactions have meaningful synergistic covariate patterns in the virus antigenicity.

Keywords

High-Order Interaction; Heredity Structure; Hierarchical Sparsity

1. INTRODUCTION

Fitting a linear regression model to the response based on a number of covariates (features) is a commonly used tool in statistical analysis. However, in numerous situations, a linear model on the covariates may be not sufficiently enough to provide comprehensive explanations for the data and to make accurate predictions. For example, in the influenza antigenic sites identification problem, a mutation at an individual antigenic site (covariate) is less deterministic to change the phenotypic behavior (antigenic change) of the influenza virus. Instead, multiple simultaneous mutations at different antigenic sites will significantly enhance the antigenic drift, and strong interactions among the antigenic sites are observed during the virus evolution.

Actually, recent statistical results have shown that studying the feature interactions in a learning model can significantly enhance its interpretability for the data and improve the prediction accuracy [6, 15, 13]. Generally, the interaction effects are represented as the elementwise product among the covariates and for example, the second-order interaction between two covariates $\mathbf{x}_i$ and $\mathbf{x}_j$ is represented by their elementwise product $\mathbf{x}_i \odot \mathbf{x}_j$. Hence, the interactions can encourage capturing the nonlinearity in the data. The interactions among covariates have been found to play an important role in various areas. For example, strong evidences have been found in [2] that the genetic-environmental interactions have significant effects on conduct disorders, and similar results are reported in [5] that the genetic environmental interactions in serotonin system are highly correlated with the adolescent depression. Moreover, in [19], considering the interaction between the continuance commitment and affective commitment is shown to be effective in predicting the absenteeism. Recently, in the antigenic sites identification problem [23], interactions among co-evolved antigenic sites are proved to be critical to quantify the impact of multiple simultaneous mutations.

As the study of the interactions among covariates gains increasing attentions, a major challenge is the exponentially expanded feature space. That is, when considering the kth-order interactions among covariates, the number of interactions is $O(d^k)$ with respect to the d covariates. Such a large number of interactions make the learning model computationally demanding even when d and k are very small. One promising strategy is to exploit sparse structure under this scenario, since only a subset of the covariates and the interactions could be of interest. A simple way is to directly apply the Lasso [21] method by treating all the covariates and the interactions equally, which is referred to the all-pairs Lasso [1, 23]. Furthermore, since the interaction effects are generated from the covariates and the higher-order interaction effects originate from lower-order ones, logical heredity relationship among those effects could be taken into account instead of treating them equally.

In order to make use of the heredity structure, statisticians favor the sparsity which obeys certain logical heredity constraints,

KDD '16, August 13-17, 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4232-2/16/08...$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939786

referring to the situation that if a set of parameters are estimated as zeros, then the set of its dependent parameters based on some certain heredity relationship should also be set to zeros. Accordingly, a number of Hierarchical Sparse Models (HSM) have been developed. For example, in [15], a convex Lasso-style method named VANISH is proposed by enforcing the *strong heredity* constraint for the second-order interactions that if a second-order interaction is added to the model, then both the corresponding covariates must be included as well. Consequently, many convex formulations, including the glinternet [11], GRESH [16], FAMILY [9], and the hierarchical sparse model [22], incorporate the strong heredity into the second-order interactions with the heredity structure achieved via the group sparsity [25], where the covariate and interactions restricted by a heredity constraint form a group. Similar considerations are discussed by [26, 24]. On the other hand, in contrast to those convex models, the SHIM method [3] adopts a non-convex formulation to achieve the strong heredity by decomposing the coefficient of each interaction into a product of the coefficients for the covariates. In addition to the strong heredity, there is another type of hierarchical relation, the *weak heredity*, which introduces a constraint that a pairwise interaction is considered if either of its corresponding covariates was included. Both the strong and weak heredity are investigated in [1] and an efficient algorithm to handle the weak heredity is introduced in [12].

So far, many interests have been focused on exploring the sparse heredity in the interaction model but none of them can deal with general hierarchies, since all of them study only the second-order interactions and their algorithms are particularly designed for the second-order interactions. On the other hand, there have been sufficient evidences to indicate that higher-order interactions are more important in many applications. For example, in psychological analysis, the third-order interactions among covariates have been shown to be important [4]. Specifically, in antigenicity analysis of influenza virus, a recent study on proteins of the H3N2 influenza virus shows that more than two of the amino acid positions could mutate simultaneously [17] and biological evidences in [14] also demonstrate that the co-evolved antigenic sites are more likely to be physically close in the 3D structure of the protein.

Unfortunately, due to the difficulty in defining and learning with the high-order heredity, we are unaware of any existing work that can deal with general hierarchies with the order of feature interactions larger than two and there is even no formal definition for the arbitrary-order heredity. In this paper, we propose a Generalized Hierarchical Sparse Model (GHSM) to tackle arbitrary-order interactions among features. We first introduce the definition of the *arbitrary-order heredity*, which makes an assumption that given any covariate, if none of its associated kth-order interaction effects contribute to a learning model, then neither do its associated higher-order interaction effects. Based on this definition, we formulate the GHSM model by applying the ℓ_1 penalty to all the coefficients under certain hierarchical chain constraints, which guarantee the arbitrary-order heredity. The resulting problem is non-convex and not easy to be optimized since the number of variables in the optimization problem increases dramatically when the order of interactions becomes bigger, which poses a computational challenge. To optimize the objective function, we use the GIST method [7] where the proximal operator is solved by the ADMM method. In the three subproblems of the ADMM method, the first two need to solve quadratic programming problems and the last one is a least square problem with a hierarchical chain constraint. After analysis, we show that all the three subproblems admit efficiently analytical solutions. In the experiments, we evaluate the GHSM method in both synthetic problem and the antigenic sites identification problem in influenza virus data, and empirical results show that the GHSM method can capture meaningful synergistic covariate patterns, which can be well explained by biological knowledge.

2. PRELIMINARIES

Throughout this paper, we use regular letters to denote scalars, bold-face and lowercase letters for vectors, and bold-face and uppercase letters for matrices or tensors. Suppose the data matrix for training is denoted by $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_d) \in \mathbb{R}^{n\times d}$, where n is the number of samples, d is the feature dimensionality, and $\mathbf{x}_i$ records the values for the ith covariate in the n data samples. The response vector is $\mathbf{y} \in \mathbb{R}^n$. The second-order interaction models [15, 3, 1, 11, 12, 16, 9, 22] commonly consider the following regression model:

$$\mathbf{y} = \sum_{i=1}^{d} \beta_i \mathbf{x}_i + \sum_{i\neq j}^{d} \phi_{i,j}(\mathbf{x}_i \odot \mathbf{x}_j) + \boldsymbol{\varepsilon}, \tag{1}$$

where $\odot$ denotes the elementwise product between vectors, $\boldsymbol{\beta} \in \mathbb{R}^d$ with β_i as its ith element is the coefficient vector for the covariates, $\boldsymbol{\Phi} \in \mathbb{R}^{d\times d}$ with $\phi_{i,j}$ as its (i,j)th element is the coefficient matrix for the pairwise interaction effects, and $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is a Gaussian noise vector. In the existing works, two types of heredity structure are considered for the second-order interactions, i.e. the strong heredity and the weak heredity, whose definitions are as follows:

$$\begin{aligned}
\text{strong heredity: } & \phi_{i,j} \neq 0 \Longrightarrow \beta_i \neq 0 \text{ and } \beta_j \neq 0, && (2)\\
& \text{or equivalently,} \\
& \beta_i = 0 \text{ or } \beta_j = 0 \Longrightarrow \phi_{i,j} = 0; \\
\text{weak heredity: } & \phi_{i,j} \neq 0 \Longrightarrow \beta_i \neq 0 \text{ or } \beta_j \neq 0, && (3)\\
& \text{or equivalently,} \\
& \beta_i = 0 \text{ and } \beta_j = 0 \Longrightarrow \phi_{i,j} = 0.
\end{aligned}$$

Based on Eqs. (2) and (3), the HSM methods introduced in [1, 12], named as the strong and weak hierNet, explicitly enforce the heredity structure by adding inequality or symmetry constraints to the Lasso method as

$$\begin{aligned}
\text{strong hierNet: } & \min_{\boldsymbol{\beta},\boldsymbol{\Phi}} l(\boldsymbol{\beta}, \boldsymbol{\Phi}|\mathbf{X}, \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\Phi}\|_1 && (4)\\
& \text{s.t. } \boldsymbol{\Phi} = \boldsymbol{\Phi}^\top,\ |\beta_i| \geq \|\boldsymbol{\Phi}_{i,\cdot}\|_1\ \forall i \in \mathbb{N}_d; \\
\text{weak hierNet: } & \min_{\boldsymbol{\beta},\boldsymbol{\Phi}} l(\boldsymbol{\beta}, \boldsymbol{\Phi}|\mathbf{X}, \mathbf{y}) + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\lambda}{2}\|\boldsymbol{\Phi}\|_1 && (5)\\
& \text{s.t. } |\beta_i| \geq \|\boldsymbol{\Phi}_{i,\cdot}\|_1\ \forall i \in \mathbb{N}_d,
\end{aligned}$$

where $l(\cdot)$ is a loss function based on Eq. (1), λ is a regularization parameter that controls the sparsity, $\mathbb{N}_d$ denotes the set of integers $\{1, \cdots, d\}$, $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector or matrix, and $\boldsymbol{\Phi}_{i,\cdot}$ denotes the ith row of $\boldsymbol{\Phi}$. The only difference between problems (4) and (5) is the existence of the symmetry constraint on $\boldsymbol{\Phi}$. It is not hard to see that the constraints in problems (4) and (5) can guarantee the strong and weak heredity defined in Eqs. (2) and (3). The strong and weak hierNet methods are representatives of the second-order HSM methods which are explicitly enforced to obey the heredity structure.

In the hierNet models, only the second-order interaction is considered. As discussed in the previous section, higher-order interactions are important to model the biological data. To the best of our knowledge, there is no work to even define the high-order interactions. In the next section, we will first provide a formal definition of the arbitrary-order heredity and then introduce our method to model the arbitrary-order interactions.

3. THE GHSM

Here we consider up to the Kth-order interactions among the covariates ($K \ll d$), and the regression model is formulated as

$$\mathbf{y} = \sum_{i=1}^{d} \theta_i^{(1)} \mathbf{z}_i^{(1)} + \sum_{i_1 < i_2}^{d} \theta_{\langle i_1, i_2 \rangle}^{(2)} \mathbf{z}_{\langle i_1, i_2 \rangle}^{(2)} + \cdots \tag{6}$$
$$+ \sum_{i_1 < i_2 < \cdots < i_K}^{d} \theta_{\langle i_1, i_2, \cdots, i_K \rangle}^{(K)} \mathbf{z}_{\langle i_1, i_2, \cdots, i_K \rangle}^{(K)} + \boldsymbol{\varepsilon},$$

where $\mathbf{z}_i^{(1)} = \mathbf{x}_i$, $\mathbf{z}_{\langle i_1, i_2, \cdots, i_k \rangle}^{(k)} = \mathbf{x}_{i_1} \odot \mathbf{x}_{i_2} \odot \cdots \odot \mathbf{x}_{i_k}$ denotes a data vector for the kth-order interaction corresponding to $\langle i_1, \cdots, i_k \rangle$, an interaction index $\langle i_1, \cdots, i_k \rangle$, where $i_1 < \cdots < i_k$, is an index to uniquely indicate the interaction among the covariates $i_1, \cdots, i_k$, $\boldsymbol{\theta}^{(k)} \in \mathbb{R}^{\binom{d}{k}}$ for $k = 1, \cdots, K$ is a vector of length $\binom{d}{k} = \frac{d!}{k!(d-k)!}$ with $\theta_{\langle i_1, \cdots, i_k \rangle}^{(k)}$ as its element corresponding to the index $\langle i_1, \cdots, i_k \rangle$. In Eq. (6), each interaction term only corresponds to one model coefficient and hence the number of model parameters is reduced from $O\left(\sum_{k=1}^{K} d^k\right)$ to $O\left(\sum_{k=1}^{K} \binom{d}{k}\right)$.

3.1 Arbitrary-Order Heredity

Based on the regression model in Eq. (6), we will define the arbitrary-order heredity in this section. We first introduce some notations. For $i_1 < \cdots < i_k$ and $j \notin \{i_1, \cdots, i_k\}$, $\langle i_1, \cdots, i_k \rangle \cup j$ and $j \cup \langle i_1, \cdots, i_k \rangle$ are the indices for the interaction effect among covariates $i_1, \cdots, i_k$ and j, which adds j into $\langle i_1, \cdots, i_k \rangle$ by preserving the ascending order. Similarly, $\langle i_1, \cdots, i_k \rangle \backslash j$ defines the index by removing the element j from $\langle i_1, \cdots, i_k \rangle$, where $j \in \{i_1, \cdots, i_k\}$.

If we follow the concept of the strong and weak heredity for the second-order case in Eqs. (2) and (3), a straightforward definition for the strong arbitrary-order heredity can be formulated as

$$\theta_{\langle i_1, \cdots, i_k \rangle}^{(k)} \neq 0 \Longrightarrow \forall j \in \{i_1, \cdots, i_k\},\ \theta_{\langle i_1, \cdots, i_k \rangle \backslash j}^{(k-1)} \neq 0,\ (k \geq 2),$$

or equivalently,

$$\theta_{\langle i_1, \cdots, i_k \rangle}^{(k)} = 0 \Longrightarrow \forall j \notin \{i_1, \cdots, i_k\},\ \theta_{\langle i_1, \cdots, i_k \rangle \cup j}^{(k+1)} = 0, (k \geq 1).$$

Similar extension can be made for the weak heredity for arbitrary-order case as well. Unfortunately, the above definition will lead to $\sum_{k=1}^{K-1} \binom{d}{k}$ constraints if the interactions up to the Kth-order are considered, which makes the problem intractable to be solved.

So, in order to represent the heredity structure in arbitrary-order case, we propose a more concise and intuitive definition as follows. First we define an ordering of the elements in $\boldsymbol{\theta}^{(k)}$ in a way following the index principle in Table 1. That is, the 1st element in $\boldsymbol{\theta}^{(k)}$ indicates the element with the interaction index $\langle 1, \cdots, k-1, k \rangle$, the 2nd element in $\boldsymbol{\theta}^{(k)}$ indicates the one with the interaction index $\langle 1, \cdots, k-1, k+1 \rangle$, and so on. Based on this ordering, we define $\mathbf{e}_i^{(k)} \in \mathbb{R}^{\binom{d}{k}}$ as a 0/1 binary vector for the ith covariate where if i appears in an interaction index of $\boldsymbol{\theta}^{(k)}$, then the corresponding element in $\mathbf{e}_i^{(k)}$ is set to 1 while the rest entries in $\mathbf{e}_i^{(k)}$ are 0.

Then we introduce the definition for the arbitrary-order heredity.

Definition 1 (Arbitrary-Order Heredity). *Given the regression model in Eq. (6), when up to the Kth-order interactions among the covariates ($K \geq 2$) are considered, the heredity among the K orders is defined as*

$$\mathbf{e}_i^{(k)} \odot \boldsymbol{\theta}^{(k)} = \mathbf{0} \Longrightarrow \mathbf{e}_i^{(k+1)} \odot \boldsymbol{\theta}^{(k+1)} = \mathbf{0},\ \forall i \in \mathbb{N}_d,\ k \in \mathbb{N}_{K-1}.$$

In Definition 1, for any covariate i, if none of its associated kth-order ($k < K$) interaction terms contribute to the regression model, then neither do its associated higher-order interaction terms. The arbitrary-order heredity poses constraints on sets of variables associated with each covariate i instead of each individual interaction coefficient and hence it only leads to $(K-1)d$ constraints, whose size is much smaller than $\sum_{k=1}^{K-1} \binom{d}{k}$, which is the size of constraints induced by the straightforward strong arbitrary-order heredity discussed before, making the learning model have much lower complexity as we will see later.

Table 1: The ordering of elements in $\boldsymbol{\theta}^{(k)}$.

The index in $\boldsymbol{\theta}^{(k)}$	The corresponding interaction index
1	$\langle 1, \cdots, k-1, k \rangle$
2	$\langle 1, \cdots, k-1, k+1 \rangle$
$\cdots$	$\cdots$
$d-k+1$	$\langle 1, \cdots, k-2, k-1, d \rangle$
$d-k+2$	$\langle 1, \cdots, k-2, k, k+1 \rangle$
$\cdots$	$\cdots$
$\binom{d}{k}$	$\langle d-k+1, \cdots, d-1, d \rangle$

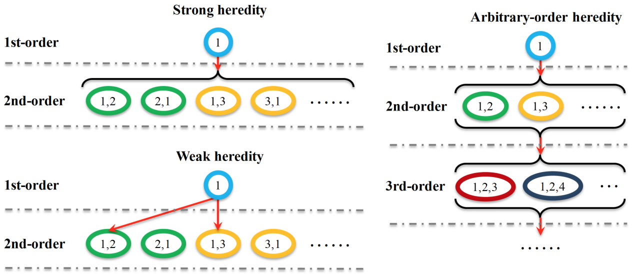


Figure 1: A pictorial illustration of the relationship among the strong, weak and arbitrary-order heredity.

When K is set to 2, it is easy to see that the arbitrary-order heredity in Definition 1 degenerates to the strong heredity in Eq. (2) and hence the arbitrary-order heredity is a generalization of the strong second-order heredity. Fig. 1 gives an example to explain the relationship among the strong, weak and arbitrary-order heredity where the circle denotes the coefficient of a covariate '1', ovals denote the coefficients of the interactions involving the covariates '1', and the red arrows indicate the heredity. From the figure, we see that when $K = 2$, the top 2 layers in the arbitrary-order heredity degenerates to the strong heredity structure by eliminating the redundancy among the model coefficients.

3.2 The Model

Based on Eq. (6) and Definition 1, we formulate the GHSM model for up to the Kth-order interactions as

$$\min_{\boldsymbol{\Theta}}\ L(\boldsymbol{\Theta}) + \sum_{k=1}^{K} \frac{\lambda}{\alpha^k} \|\boldsymbol{\theta}^{(k)}\|_1, \tag{7}$$
$$\text{s.t. } |\theta_i^{(1)}| \geq \|\mathbf{e}_i^{(2)} \odot \boldsymbol{\theta}^{(2)}\|_1 \geq \cdots \geq \|\mathbf{e}_i^{(K)} \odot \boldsymbol{\theta}^{(K)}\|_1,\ i \in \mathbb{N}_d,$$

where λ and α are two regularization parameters controlling the sparsity and the decay in the coefficients for interactions of different orders, $\boldsymbol{\Theta}$ denotes the set of parameters $\{\boldsymbol{\theta}^{(k)}\}_{k=1}^{K}$, and $L(\cdot)$ is a loss function for regression such as the square loss defined as $L(\boldsymbol{\Theta}) = \frac{1}{2}\|\mathbf{y} - \sum_{k=1}^{K} \sum_{i_1 < \cdots < i_k}^{d} \theta_{\langle i_1, \cdots, i_k \rangle}^{(k)} \mathbf{z}_{\langle i_1, \cdots, i_k \rangle}^{(k)}\|_2^2$ where $\|\cdot\|_2$ denotes the ℓ_2 norm of a vector. In problem (7), the constraints associated with each covariate i have a chain of inequality constraints, which contains $(K-1)$ inequality constraints and there are a total of d chains. It is easy to see that these constraints achieve the arbitrary-order hierarchy in Definition 1.

Problem (7) is non-convex due to the chains of inequality constraints. Moreover, the variables are coupled in different chains of constraints, e.g., the variable $\theta^{(3)}_{\langle 1,2,3\rangle}$ appears in three chains associated with the covariates $1, 2, 3$ respectively, which makes the problem more complex. In the next section, we propose an efficient algorithm to solve problem (7).

4. OPTIMIZATION ALGORITHM

In this section, we introduce the optimization algorithm to solve problem (7). The main idea is to combine proximal gradient methods and the ADMM. Since problem (7) is non-convex, we adopt the GIST method [7] whose entire algorithm is shown in Algorithm 1. We define $r(\boldsymbol{\Theta})$ as

$$r(\boldsymbol{\theta}) = \begin{cases} \sum_{k=1}^{K} \frac{\lambda}{\alpha^k} \|\boldsymbol{\theta}^{(k)}\|_1, & \text{if the constraint in Eq. (7) is satisfied;} \\ +\infty, & \text{otherwise.} \end{cases}$$

Then, the proximal operator at the $(t+1)$th iteration in the GIST method solves the following problem:

$$\boldsymbol{\Theta}^{(t+1)} = \arg\min_{\boldsymbol{\Theta}} \frac{\tau_t}{2} \sum_{k=1}^{K} \|\boldsymbol{\theta}^{(k)} - \mathbf{v}^{(k)}\|_2^2 + r(\boldsymbol{\Theta}), \tag{8}$$

where $[\boldsymbol{\theta}^{(k)}]^{(t)}$ denotes the estimation of $\boldsymbol{\theta}^{(k)}$ in the tth iteration, $\mathbf{v}^{(k)} = [\boldsymbol{\theta}^{(k)}]^{(t)} - \frac{1}{\tau_t} \nabla_{\boldsymbol{\theta}^{(k)}} L(\boldsymbol{\Theta}^{(t)})$, $\nabla_{\boldsymbol{\theta}^{(k)}} L(\boldsymbol{\Theta}^{(t)})$ denotes the gradient of $L(\boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\theta}^{(k)}$ at $\boldsymbol{\Theta}^{(t)}$ and it can be easily computed as $[\mathbf{Z}^{(k)}]^\top (\sum_{k'=1}^{K} \mathbf{Z}^{(k')} \boldsymbol{\theta}^{(k')} - \mathbf{y})$ for $k \in \mathbb{N}_K$, $\mathbf{Z}^{(k)} \in \mathbb{R}^{n \times \binom{d}{k}}$ is the matrix containing all the kth-order interaction effects, and τ_t is a step size determined via a line search method by satisfying the following condition:

$$F(\boldsymbol{\Theta}^{(t+1)}) \le F(\boldsymbol{\Theta}^{(t)}) - \frac{\varphi \tau_t}{2} \sum_{k=1}^{K} \|[\boldsymbol{\theta}^{(k)}]^{(t+1)} - [\boldsymbol{\theta}^{(k)}]^{(t)}\|_2^2, \tag{9}$$

where $F(\boldsymbol{\Theta}) = L(\boldsymbol{\Theta}) + r(\boldsymbol{\Theta})$ and φ is a constant in $(0, 1)$. Then, the GIST algorithm iteratively solves the proximal problem (8) until convergence.

Algorithm 1 The GIST algorithm for solving problem (7).

```
Input: $\mathbf{X}, \mathbf{y}, K, \epsilon = 10^{-4}$;
Output: $\hat{\boldsymbol{\Theta}}$;
Initialize $\boldsymbol{\Theta}^{(0)}, \eta > 1, 0 < \tau_{min} < \tau_{max}, \varphi \in (0,1), t = 0$;
repeat
    $\tau_t \in [\tau_{min}, \tau_{max}]$;
    repeat
        Solve the proximal problem (8) with $\boldsymbol{\Theta}^{(t)}$ and $\tau_t$;
        $\tau_t = \eta \tau_t$;
    until condition (9) is satisifed;
    $t = t + 1$;
until $F(\boldsymbol{\Theta}^{(t)}) - F(\boldsymbol{\Theta}^{(t+1)}) < \epsilon$;
$\hat{\boldsymbol{\Theta}} = \boldsymbol{\Theta}^{(t)}$;
```

4.1 Solving The Proximal Problem

The key problem in Algorithm 1 is to solve the proximal problem (8). Since $r(\cdot)$ is an extended real-value function, problem (8) can be reformulated as

$$\min_{\boldsymbol{\Theta}} \frac{\tau}{2} \sum_{k=1}^{K} \|\boldsymbol{\theta}^{(k)} - \mathbf{v}^{(k)}\|_2^2 + \sum_{k=1}^{K} \frac{\lambda}{\alpha^k} \|\boldsymbol{\theta}^{(k)}\|_1, \tag{10}$$
$$\text{s.t. } |\theta_i^{(1)}| \ge \|\mathbf{e}_i^{(2)} \odot \boldsymbol{\theta}^{(2)}\|_1 \ge \cdots \ge \|\mathbf{e}_i^{(K)} \odot \boldsymbol{\theta}^{(K)}\|_1 \ \forall i \in \mathbb{N}_d,$$

where we omit the iterative index t for notational simplicity. Problem (10) is still non-convex due to the chains of inequality constraints. However, the following theorem shows that problem (10) admits an equivalently convex formulation.[1]

THEOREM 1. *Let $\bar{\boldsymbol{\theta}}^{(k)} = |\boldsymbol{\theta}^{(k)}|$, where the operator $|\cdot|$ denotes the elementwise absolute operator on a vector. Then problem (10) is equivalent with the following convex optimization problem:*

$$\min_{\bar{\boldsymbol{\theta}}^{(1)}, \cdots, \bar{\boldsymbol{\theta}}^{(K)}} \frac{\tau}{2} \sum_{k=1}^{K} \|\bar{\boldsymbol{\theta}}^{(k)} - \bar{\mathbf{v}}^{(k)}\|_2^2, \tag{11}$$
$$\text{s.t. } \bar{\theta}_i^{(1)} \ge [\mathbf{e}_i^{(2)}]^\top \bar{\boldsymbol{\theta}}^{(2)} \ge \cdots \ge [\mathbf{e}_i^{(K)}]^\top \bar{\boldsymbol{\theta}}^{(K)}, \ \forall i \in \mathbb{N}_d,$$
$$\bar{\boldsymbol{\theta}}^{(k)} \succeq 0, \ \forall k \in \mathbb{N}_K,$$

where $\mathbf{1}$ denotes a column vector of all ones with appropriate size, $\bar{\mathbf{v}}^{(k)} = |\mathbf{v}^{(k)}| - \frac{\lambda}{\alpha^k \tau} \mathbf{1}$, and $\succeq$ denotes the elementwise 'no smaller than' operator. The solution of problem (10) can be obtained as $\boldsymbol{\theta}^{(k)} = \operatorname{sign}(\mathbf{v}^{(k)}) \odot \bar{\boldsymbol{\theta}}^{(k)}$ for $k \in \mathbb{N}_K$, where $\operatorname{sign}(\cdot)$ is the elementwise sign operator.

It is easy to find that problem (11) is a quadratic programming problem, hence many off-the-shelf solvers for convex programming can be used directly to obtain the optimal solution. Nevertheless, instead of using these tools, we propose an efficient algorithm to solve problem (11) by taking advantage of the chain structure in the constraints. In problem (11), the variables are coupled together in the chains of inequality constraints. In order to decouple these parameters, we use the ADMM method to solve problem (11) by introducing new variables.

We define $\mathbf{p}^{(k)} = \bar{\boldsymbol{\theta}}^{(k)}$ for $k \in \mathbb{N}_K$, $\boldsymbol{\delta}_i = ([\mathbf{e}_i^{(2)}]^\top \mathbf{p}^{(2)}, [\mathbf{e}_i^{(3)}]^\top \mathbf{p}^{(3)}, \cdots, [\mathbf{e}_i^{(K)}]^\top \mathbf{p}^{(K)})^\top \in \mathbb{R}^{K-1}$ for $i \in \mathbb{N}_d$, and $\mathbf{q}_i = \boldsymbol{\delta}_i \in \mathbb{R}^{K-1}$. Then, problem (11) can be reformulated as

$$\min_{\{\bar{\boldsymbol{\theta}}^{(k)}\}, \{\mathbf{p}^{(k)}\}, \{\mathbf{q}_i\}} \quad \frac{\tau}{2} \sum_{k=1}^{K} \|\bar{\boldsymbol{\theta}}^{(k)} - \bar{\mathbf{v}}^{(k)}\|_2^2, \tag{12}$$
$$\begin{aligned} \text{s.t.} \quad & \bar{\boldsymbol{\theta}}^{(k)} \succeq 0, \ \forall k \in \mathbb{N}_K \\ & \mathbf{p}^{(k)} = \bar{\boldsymbol{\theta}}^{(k)}, \ \forall k \in \mathbb{N}_K \\ & \mathbf{q}_i = \boldsymbol{\delta}_i, \ \forall i \in \mathbb{N}_d \\ & p_i^{(1)} \ge q_{1,i} \ge \cdots \ge q_{K-1,i}, \ \forall i \in \mathbb{N}_d, \end{aligned}$$

where $p_i^{(1)}$ is the ith element in $\mathbf{p}^{(1)}$ and $q_{j,i}$ is the jth element in $\mathbf{q}_i$. Based on problem (12), we define the augmented Lagrangian function as

$$\bar{L}(\bar{\boldsymbol{\Theta}}, \mathbf{P}, \mathbf{Q}) = \frac{\tau}{2} \sum_{k=1}^{K} \|\bar{\boldsymbol{\theta}}^{(k)} - \bar{\mathbf{v}}^{(k)}\|_2^2 + \sum_{k=1}^{K} \frac{\rho_1}{2} \|\mathbf{p}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}\|_2^2$$
$$+ \sum_{k=1}^{K} [\boldsymbol{a}^{(k)}]^\top (\mathbf{p}^{(k)} - \bar{\boldsymbol{\theta}}^{(k)}) + \sum_{i=1}^{d} \frac{\rho_2}{2} \|\mathbf{q}_i - \boldsymbol{\delta}_i\|_2^2 + \sum_{i=1}^{d} \boldsymbol{b}_i^\top (\mathbf{q}_i - \boldsymbol{\delta}_i).$$

where $\bar{\boldsymbol{\Theta}}$ denotes the set of parameters $\{\bar{\boldsymbol{\theta}}^{(k)}\}_{k=1}^{K}$, $\mathbf{P}$ and $\mathbf{Q}$ denote the sets of variables $\{\mathbf{p}^{(k)}\}_{k=1}^{K}$ and $\{\mathbf{q}_i\}_{i=1}^{d}$ respectively, $\{\boldsymbol{a}^{(k)}\}_{k=1}^{K}$ and $\{\boldsymbol{b}_i\}_{i=1}^{d}$ are the Lagrangian multipliers, and ρ_1 and ρ_2 are two penalty parameters. Then we need to solve the following problem:

$$\min_{\bar{\boldsymbol{\Theta}}, \mathbf{P}, \mathbf{Q}} \bar{L}(\bar{\boldsymbol{\Theta}}, \mathbf{P}, \mathbf{Q}) \text{ s.t. } \begin{cases} \bar{\boldsymbol{\theta}}^{(k)} \succeq 0, \ k \in \mathbb{N}_K; \\ p_i^{(1)} \ge q_{1,i} \ge \cdots \ge q_{K-1,i}, \ i \in \mathbb{N}_d. \end{cases}$$

The above problem can be solved via the ADMM algorithm presented in algorithm 2, in which three subproblems in steps 4, 5 and 6 need to be solved. In the next two sections, we will show how to solve those subproblems efficiently.

[1] We put all the proofs in the supplementary material at http://www.stat.rutgers.edu/home/lhan.

Algorithm 2 The ADMM algorithm for solving problem (12).

```
Input: $\mathbf{X}, \mathbf{y}, K$;
Output: $\hat{\mathbf{\Theta}}$;
 1: Initialize $\bar{\mathbf{\Theta}}^{(0)}$, $\mathbf{Q}^{(0)}$ and $\mathbf{A}^{(0)}$;
 2: Set $\rho = 0.1$ and $t = 0$;
 3: repeat
 4:     Solve $\bar{\mathbf{\Theta}}^{(t+1)}$ with fixed $\mathbf{P}^{(t)}$ and $\mathbf{Q}^{(t)}$;
 5:     Solve $\{\mathbf{P}^{(t+1)} \backslash [\mathbf{p}^{(1)}]^{(t+1)}\}$ with fixed $\bar{\mathbf{\Theta}}^{(t)}$, $\mathbf{Q}^{(t)}$ and $[\mathbf{p}^{(1)}]^{(t)}$;
 6:     Solve $\mathbf{Q}^{(t+1)}$ and $[\mathbf{p}^{(1)}]^{(t+1)}$ with fixed $\bar{\mathbf{\Theta}}^{(t)}$, $\{\mathbf{P}^{(t)} \backslash [\mathbf{p}^{(1)}]^{(t)}\}$;
 7:     $[\boldsymbol{a}^{(k)}]^{(t+1)} = [\boldsymbol{a}^{(k)}]^{(t)} + \rho_1([\mathbf{p}^{(k)}]^{(t)} - [\bar{\boldsymbol{\theta}}^{(k)}]^{(t)})$ for $k \in \mathbb{N}_K$;
 8:     $\boldsymbol{b}_i^{(t+1)} = \boldsymbol{b}_i^{(t)} + \rho_2(\mathbf{q}_i^{(t)} - \boldsymbol{\delta}_i^{(t)})$ for $i \in \mathbb{N}_d$;
 9:     $t = t + 1$;
10: until Some convergence criterion is satisfied;
```

4.2 Analytical Solutions in Steps 4 and 5

In step 4 of Algorithm 2, with fixed $\mathbf{P}$ and $\mathbf{Q}$, the problem with respect to $\bar{\mathbf{\Theta}}$ can be rewritten as

$$\min_{\bar{\mathbf{\Theta}}} \sum_{k=1}^{K} \|\bar{\boldsymbol{\theta}}^{(k)} - \frac{\tau \bar{\mathbf{v}}^{(k)} + \rho_1 \mathbf{p}^{(k)} + \boldsymbol{a}^{(k)}}{\tau + \rho_1}\|_2^2, \text{ s.t. } \bar{\boldsymbol{\theta}}^{(k)} \succeq 0, \quad (13)$$

which can be easily solved via the following analytical solution:

$$[\bar{\boldsymbol{\theta}}^{(k)}]^* = \max\left(0, \frac{\tau \bar{\mathbf{v}}^{(k)} + \rho_1 \mathbf{p}^{(k)} + \boldsymbol{a}^{(k)}}{\tau + \rho_1}\right), \forall k \in \mathbb{N}_K. \quad (14)$$

Given fixed $\bar{\mathbf{\Theta}}$, $\mathbf{Q}$ and $\mathbf{p}^{(1)}$, the problem with respect to $\mathbf{p}^{(2)}, \cdots, \mathbf{p}^{(K)}$ corresponding to step 5 of Algorithm 2 can be decomposed into $K-1$ separable problems with the problem for $\mathbf{p}^{(k)}$ formulated as

$$\min_{\mathbf{p}^{(k)}} \frac{1}{2}[\mathbf{p}^{(k)}]^\top \mathbf{H}^{(k)} \mathbf{p}^{(k)} - [\mathbf{c}^{(k)}]^\top \mathbf{p}^{(k)}, \ k = 2, \cdots, K, \quad (15)$$

where $\mathbf{I}$ is an identity matrix with appropriate size, the matrix $\mathbf{E}^{(k)} = (\mathbf{e}_1^{(k)}, \mathbf{e}_2^{(k)}, \cdots, \mathbf{e}_d^{(k)}) \in \mathbb{R}^{\binom{d}{k} \times d}$, and $\mathbf{H}^{(k)}$, $\mathbf{c}^{(k)}$ are defined as

$$\begin{aligned} \mathbf{H}^{(k)} =& \mathbf{I} + \mathbf{E}^{(k)} [\mathbf{E}^{(k)}]^\top, \\ \mathbf{c}^{(k)} =& \bar{\boldsymbol{\theta}}^{(k)} - \frac{1}{\rho_1}\boldsymbol{a}^{(k)} + \sum_{i=1}^{d} (\frac{1}{\rho_2} b_{k-1,i} + q_{k-1,i}) \mathbf{e}_i^{(k)}. \end{aligned}$$

Note that $\mathbf{H}^{(k)}$ is positive definite (PD). Hence, $\mathbf{H}^{(k)}$ is invertible and its inverse can be computed efficiently as $[\mathbf{H}^{(k)}]^{-1} = \mathbf{I} - \mathbf{E}^{(k)}(\mathbf{I} + [\mathbf{E}^{(k)}]^\top \mathbf{E}^{(k)})^{-1} [\mathbf{E}^{(k)}]^\top$, where the matrix inverse is actually taken on a $d \times d$ matrix $\mathbf{I} + [\mathbf{E}^{(k)}]^\top \mathbf{E}^{(k)}$ instead of the original $\binom{d}{k} \times \binom{d}{k}$ matrix $\mathbf{H}^{(k)}$. Then, the optimal solution of problem (15) can be computed as

$$\begin{aligned} [\mathbf{p}^{(k)}]^* &= [\mathbf{H}^{(k)}]^{-1} \mathbf{c}^{(k)} \quad (16) \\ &= \mathbf{c}^{(k)} - \left(\mathbf{E}^{(k)} (\mathbf{I} + [\mathbf{E}^{(k)}]^\top \mathbf{E}^{(k)})^{-1}\right) \left([\mathbf{E}^{(k)}]^\top \mathbf{c}^{(k)}\right). \end{aligned}$$

Moreover, given the data, $\mathbf{E}^{(k)}(\mathbf{I} + [\mathbf{E}^{(k)}]^\top \mathbf{E}^{(k)})^{-1} \in \mathbb{R}^{\binom{d}{k} \times d}$ is a constant matrix and so it can be computed only once and stored prior to the model learning. As a consequence, the analytical solution in Eq. (16) only takes $O(d\binom{d}{k})$ time complexity for each $[\bar{\boldsymbol{\theta}}^{(k)}]^*$, which is almost *linear* with respect to the number of the kth-order interactions.

4.3 Efficient Solution in Step 6

With fixed $\bar{\mathbf{\Theta}}$ and $\mathbf{p}^{(2)}, \cdots, \mathbf{p}^{(K)}$, the problem w.r.t. $\mathbf{Q}$ and $\mathbf{p}^{(1)}$ can be formulated as

$$\begin{aligned} \min_{\mathbf{Q}, \mathbf{p}^{(1)}} & \frac{\rho_1}{2} \|\mathbf{p}^{(1)} - (\bar{\boldsymbol{\theta}}^{(1)} - \frac{\boldsymbol{a}^{(1)}}{\rho_1})\|_2^2 + \frac{\rho_2}{2} \sum_{i=1}^{d} \|\mathbf{q}_i - (\boldsymbol{\delta}_i - \frac{\boldsymbol{b}_i}{\rho_2})\|_2^2 \\ \text{s.t. } & p_i^{(1)} \geq q_{1,i} \geq \cdots \geq q_{K-1,i}, \ \forall i \in \mathbb{N}_d. \quad (17) \end{aligned}$$

Problem (17) can be decomposed into d independent problems with the ith one formulated as

$$\begin{aligned} \min_{\mathbf{q}_i, p_i^{(1)}} & \frac{\rho_1}{\rho_2} \left(p_i^{(1)} - (\bar{\theta}_i^{(1)} - \frac{a_i^{(1)}}{\rho_1})\right)^2 + \sum_{k=1}^{K-1} \left(q_{k,i} - (\delta_{k,i} - \frac{b_{k,i}}{\rho_2})\right)^2 \\ \text{s.t. } & p_i^{(1)} \geq q_{1,i} \geq \cdots \geq q_{K-1,i}, \quad (18) \end{aligned}$$

where $\bar{\theta}_i^{(1)}$, $a_i^{(1)}$ are the ith elements in $\bar{\boldsymbol{\theta}}^{(1)}$ and $\boldsymbol{a}^{(1)}$ respectively, and $\delta_{k,i}$, $b_{k,i}$ are the kth elements in $\boldsymbol{\delta}_i$ and $\boldsymbol{b}_i$ respectively.

In problem (18), the chain of inequality constraints still exists. We first rewrite this problem into a more general formulation as

$$\min_{\mathbf{s}} \sum_{k=1}^{K} \omega_k (s_k - u_k)^2 \quad \text{s.t. } s_1 \geq s_2 \geq \cdots \geq s_K, \quad (19)$$

where problem (18) is a special case of problem (19) by setting $\mathbf{s} = (p_i^{(1)}, q_{1,i}, \cdots, q_{K-1,i})^\top \in \mathbb{R}^K$, $\mathbf{u} = (\bar{\theta}_i^{(1)} - \frac{a_i^{(1)}}{\rho_1}, \delta_{1,i} - \frac{b_{1,i}}{\rho_2}, \cdots, \delta_{K-1,i} - \frac{b_{K-1,i}}{\rho_2})^\top \in \mathbb{R}^K$ and $\boldsymbol{\omega} = (\frac{\rho_1}{\rho_2}, 1, \cdots, 1)^\top \in \mathbb{R}^K$.

In the following, we generalize our previous results in [8] to show that an efficient solution exists for problem (19). We first introduce two useful lemmas to reveal some interesting properties of problem (19).

LEMMA 1. *In problem (19), the following properties hold: (1) If $u_1 \geq u_2 \geq \cdots \geq u_K$, then the optimal solution $(s_1^*, s_2^*, \cdots, s_K^*)$ is $(u_1, u_2, \cdots, u_K)$; (2) If $u_1 \leq u_2 \leq \cdots \leq u_K$, then the optimal solution $(s_1^*, s_2^*, \cdots, s_K^*)$ is $(u^*, \cdots, u^*)|_K$, where $u^* = \frac{\sum_{k=1}^K \omega_k u_k}{\sum_{k=1}^K \omega_k}$ and $(u^*, \cdots, u^*)|_K$ denotes a sequence with K identical elements u^*.*

LEMMA 2. *For any two sets of inputs $\{(u_1, \cdots, u_l), (\omega_1, \cdots, \omega_l)\}$ and $\{(u_{l+1}, \cdots, u_n), (\omega_{l+1}, \cdots, \omega_n)\}$, which define two instances of problem (19), if the optimal solutions for them are $(\dot{u}^*, \cdots, \dot{u}^*)|_l$ and $(\ddot{u}^*, \cdots, \ddot{u}^*)|_{n-l}$ respectively, then we have: (1) If $\dot{u}^* \geq \ddot{u}^*$, the optimal solution for the problem defined by the concatenated sequence $(u_1, \cdots, u_l) \bowtie (u_{l+1}, \cdots, u_n)$ and concatenated weights $(\omega_1, \cdots, \omega_l) \bowtie (\omega_{l+1}, \cdots, \omega_n)$ is $(\dot{u}^*, \cdots, \dot{u}^*)|_l \bowtie (\ddot{u}^*, \cdots, \ddot{u}^*)|_{n-l}$; (2) Otherwise, i.e., $\dot{u}^* < \ddot{u}^*$, the optimal solution for the problem defined by the concatenated sequence is $(u^*, \cdots, u^*)|_n$, where $u^* = \frac{\sum_{i=1}^n \omega_i u_i}{\sum_{i=1}^n \omega_i}$.*

Lemma 2 implies that we can immediately obtain the solution of problem (19) defined by the input $(u_1, \cdots, u_n)$ and $(\omega_1, \cdots, \omega_n)$, if $(u_1, \cdots, u_n)$ is a concatenation from two sub-sequences and the optimal solutions corresponding to problems defined by the two sub-sequences have solutions with identical values. Therefore, based on Lemma 2, we devise Algorithm 3 to solve problem (19) with its optimality guaranteed by the following theorem.

THEOREM 2. *For problem (19) defined by $(u_1, \cdots, u_K)$ and $(\omega_1, \cdots, \omega_K)$, Algorithm 3 finds its optimal solution.*

4.4 Time Complexity

We analyze the time complexity of the whole optimization procedure for solving the GHSM model. We first discuss the time complexity of the inner most Algorithm 3. In Algorithm 3, step 1 only needs to scan the input sequence $(u_1, \cdots, u_K)$ once and thus it needs $O(K)$ time. Although there exist two loops from step 3 to step 14, the maximum number of the concatenation operations in step 8 is $M - 1$. For the concatenation operation, according to Lemma 2, it only needs to compute the weighted average of the

Algorithm 3 The algorithm for solving problem (19).

```
Input: $(u_1, \cdots, u_K)$ and $(\omega_1, \cdots, \omega_K)$;
Output: $(s_1^*, \cdots, s_K^*)$;
Scan $(u_1, \cdots, u_K)$ once to split it into $M$ non-decreasing subsequences $(\boldsymbol{t}_1, \cdots, \boldsymbol{t}_M)$ and meanwhile split $(\omega_1, \cdots, \omega_K)$ accordingly. Then calculate the solutions for the problems defined by those sub-sequences and sub-weights based on Lemma 1;
Push $\boldsymbol{t}_1$ into a stack;
for $m = 2 : M$ do
    Push $\boldsymbol{t}_m$ into the stack;
    while there are at least two sequences in the stack do
        Pop the first and second sequences from the stack and denote the solutions for their associated problems by $\ddot{u}^*$ and $\dot{u}^*$ separately;
        if $\dot{u}^* < \ddot{u}^*$ then
            Concatenate the two sequences under the second condition in Lemma 2 and then push the concatenated sequence into the stack;
        else
            Push the two sequences back into the stack without any operation;
            Break;
        end if
    end while
end for
Concatenate the solutions of the sequences in the stack from bottom to top and output the concatenated solution;
```

entries in two sequences and we just need to record the weighted average and the sum of the weights in each sequence, making each concatenation operation cost $O(1)$. Since $M \leq K$, the complexity of Algorithm 3 is $O(K)$.

In Algorithm 2, solving the subproblem in step 4 by using Eq. (14) requires $O\left(\sum_{k=1}^{K}\binom{d}{k}\right)$ time. The computation of step 5 by using the closed-form solution in Eq. (16) takes $O\left(d\sum_{k=2}^{K}\binom{d}{k}\right)$ time. The computation in step 6 needs to execute Algorithm 3 for d times, and hence the time cost is $O(dK)$. Usually, we have $K \ll d$ and hence $dK < d\sum_{k=2}^{K}\binom{d}{k}$. So the total time complexity of each iteration in Algorithm 2 is $O\left(d\sum_{k=2}^{K}\binom{d}{k}\right)$.

By assuming that Algorithms 1 and 2 need N_1 and N_2 iterations to converge respectively, the total time complexity for solving the GHSM model is $O\left(N_1 N_2 d\sum_{k=2}^{K}\binom{d}{k}\right)$, which is almost *linear* with respect to the total number of interaction effects $\sum_{k=2}^{K}\binom{d}{k}$. In our experiments, we empirically find that both N_1 and N_2 are small when convergence. Hence, the overall algorithm for solving the GHSM model is very efficient. Moreover, according to Sections 4.2 and 4.3, the problems in steps 4, 5 and 6 of the ADMM algorithm can be decomposed into a number of independent problems, which can be parallelized to further improve the efficiency.

5. EXPERIMENTS

In this section, we empirically evaluate the proposed GHSM method and compare with a large number of the state-of-the-art methods for hierarchical sparse modeling. Specifically, the competitors include (1) Lasso [21], which is the sparse model using covariates only by applying the ℓ_1 penalty on the model coefficients; (2) All-Interactions Lasso (AIL-k), which is the sparse interaction model using up to the kth-order interactions based on the ℓ_1 penalty. We concatenate all the effects together to form a new data matrix and treat it as a Lasso problem; (3) weak hierNet (w-hierNet) [1], which is the HSM method using up to the 2nd-order interaction effects with the weak heredity, i.e., solving the problem in Eq. (5). Its R package 'hierNet' is available in 'CRAN'; (4) eWHL [12], which is an efficient implementation for the w-hierNet method proposed in [1]; (5) strong hierNet (s-hierNet) [1], which is the HSM method using up to the 2nd-order interaction effects with the strong heredity, i.e., solving the problem in Eq. (4). Its R package 'hierNet' is available in 'CRAN'; (6) FAMILY [9], which is a convex HSM model using up to the 2nd-order interaction effects with the strong heredity, where the sparsity is achieved by using the group lasso. Many methods [11, 16, 9, 22] have similar ideas and we select the FAMILY method as a representative. Its R package 'FAMILY' is available in 'CRAN'; (7) GHSM-k, which is the proposed GHSM models that can deal with up to the arbitrary kth-order interaction effects.

The experimental platform is a 64-bit machine with 2.2 GHz quad-core Intel Core i7 CPU and 16 GB memory.

5.1 Synthetic Study

In this section, we conduct experiments on synthetic datasets.

5.1.1 Settings

To study different orders of interactions, we generate 3 synthetic datasets with the highest order being $K = 3, 4, 5$ respectively. In all the datasets, the number of training samples n is set to 200 and the number of covariates d is 20. Each entry in the data matrix for training, $\mathbf{X} \in \mathbb{R}^{n\times d}$, is sampled from the standard normal distribution. The kth-order interaction matrix $\mathbf{Z}^{(k)} \in \mathbb{R}^{n\times\binom{d}{k}}$ is then generated with its column indices following Table 1. All the columns of matrices $\mathbf{X}$ and $\mathbf{Z}^{(k)}$'s are normalized to have zero mean and unit variance.

In all the 3 datasets, we set the first half of the coefficients with respect to $\boldsymbol{\theta}^{(1)}$ to be 1 and the rest to be 0. For the first dataset with $K = 3$, the indices of the non-zero coefficients in the second-order interactions are set to be $\langle 1, 2\rangle$, $\langle 3, 4\rangle$, $\langle 5, 6\rangle$, $\langle 7, 8\rangle$ and $\langle 9, 10\rangle$, i.e., there are 5 non-zero coefficients in $\boldsymbol{\theta}^{(2)}$. Similarly, we set the indices of the non-zero coefficients in the third-order interactions as $\langle 1, 2, 3\rangle$, $\langle 2, 3, 4\rangle$, $\langle 4, 5, 6\rangle$, $\langle 5, 6, 7\rangle$, $\langle 7, 8, 9\rangle$ and $\langle 8, 9, 10\rangle$, i.e., 6 entries in $\boldsymbol{\theta}^{(3)}$ are non-zero. All the non-zero entries in $\boldsymbol{\theta}^{(2)}$ and $\boldsymbol{\theta}^{(3)}$ are set to 0.5. For the second dataset with $K = 4$, we keep the settings for the orders up to the third order as in the case $K = 3$ and set the indices of the non-zero coefficients in the 4th-order interactions as $\langle 1, 2, 3, 4\rangle$, $\langle 2, 3, 4, 5\rangle$, $\langle 5, 6, 7, 8\rangle$ and $\langle 6, 7, 8, 9\rangle$, leading to 4 non-zero coefficients in $\boldsymbol{\theta}^{(4)}$. All the non-zero entries in $\boldsymbol{\theta}^{(4)}$ are also set to 0.5. For the third dataset with $K = 5$, we adopt the same settings for up to the fourth order as in the case $K = 4$ and set the indices of non-zero coefficients in the 5th-order interactions as $\langle 1, 2, 3, 4, 5\rangle$ and $\langle 2, 3, 4, 5, 6\rangle$ with the corresponding entries having values of 0.5. It is easy to check that the above settings of the model coefficients satisfy the arbitrary-order heredity. Finally, the response vector $\mathbf{y}$ is constructed as $\mathbf{y} = \sum_{k=1}^{K}\mathbf{Z}^{(k)}\boldsymbol{\theta}^{(k)} + \boldsymbol{\epsilon}$ where $\mathbf{Z}^{(1)} = \mathbf{X}$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}/4)$. The statistic of the interaction effects used in the synthetic data is given in Table 3.

For all the methods in comparison, we choose their regularization parameters from a set $\{0.1, 0.3, 0.5, 1, 3, 5, 10, 30, 50\}$ by using additional 200 data samples as a validation set. For the GHSM method, there is another regularization parameter α, which is selected from a set $\{1, 2, 10\}$. To measure the performance of different methods, we use the sensitivity (Sen.) and the specificity (Spe.) [12], where non-zero entries in the corresponding coefficient vector are treated as positive and zero entries are negative, for each order of interactions to test the recovery performance on the model coefficients and use the root mean square error (RMSE) on a test set having 200 samples for each setting. For each setting, we repeat each configuration for 10 times and report the average results.

Table 2: **The recovery and prediction performance of different methods averaged over 10 repetitions on the synthetic data. '−' indicates the value is not available in the corresponding setting. Higher values of Sen. and Spe. indicates better recovery performance on non-zero and zero entries, respectively. The numbers in bold denote the best results.**

	Synthetic data 1: with up to the 3rd-order interactions							Synthetic data 2: with up to the 4th-order interactions								
	Recovery on training data							Recovery on training data								
	1st-order		2nd-order		3rd-order		Testing	1st-order		2nd-order		3rd-order		4th-order		Testing
Method	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	RMSE	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	RMSE
Lasso	**1.00**	0.86	−	−	−	−	2.910±0.283	**1.00**	0.87	−	−	−	−	−	−	3.010±0.225
AIL-2	**1.00**	0.94	**1.00**	0.92	−	−	1.773±0.244	**1.00**	0.93	**1.00**	0.90	−	−	−	−	2.006±0.308
w-hierNet	**1.00**	0.85	**1.00**	0.95	−	−	1.702±0.272	**1.00**	0.82	**1.00**	0.94	−	−	−	−	1.929±0.290
eWHL	**1.00**	0.92	**1.00**	0.94	−	−	1.718±0.241	**1.00**	0.88	**1.00**	0.95	−	−	−	−	1.907±0.273
s-hierNet	**1.00**	0.79	**1.00**	0.96	−	−	1.655±0.257	**1.00**	0.60	**1.00**	0.89	−	−	−	−	1.868±0.267
FAMILY	**1.00**	0.93	**1.00**	0.97	−	−	1.639±0.211	**1.00**	0.83	**1.00**	0.92	−	−	−	−	1.891±0.299
GHSM-2	**1.00**	0.97	**1.00**	**0.99**	−	−	1.672±0.214	**1.00**	0.88	**1.00**	**0.99**	−	−	−	−	1.856±0.249
AIL-3	**1.00**	0.93	**1.00**	0.94	0.81	0.96	1.027±0.224	**1.00**	0.88	**1.00**	0.86	0.95	0.90	−	−	1.704±0.161
GHSM-3	**1.00**	**0.98**	**1.00**	0.97	**1.00**	**0.97**	**0.965±0.246**	**1.00**	**0.97**	**1.00**	0.91	**1.00**	0.91	−	−	1.506±0.168
AIL-4	−	−	−	−	−	−	−	**1.00**	0.95	**1.00**	0.94	0.93	**0.96**	0.85	0.98	1.484±0.226
GHSM-4	−	−	−	−	−	−	−	**1.00**	0.96	0.98	0.86	0.92	0.91	**0.93**	**0.99**	**1.333±0.271**

	Synthetic data 3: with up to the 5th-order interactions										
	Recovery on training data										
	1st-order		2nd-order		3rd-order		4th-order		5th-order		Testing
Method	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	Sen.	Spe.	RMSE
Lasso	**1.00**	0.85	−	−	−	−	−	−	−	−	3.239±0.366
AIL-2	**1.00**	**0.95**	**1.00**	0.88	−	−	−	−	−	−	2.279±0.398
w-hierNet	**1.00**	0.83	**1.00**	0.94	−	−	−	−	−	−	2.157±0.388
eWHL	**1.00**	0.90	**1.00**	**0.95**	−	−	−	−	−	−	2.111±0.372
s-hierNet	**1.00**	0.46	**1.00**	0.89	−	−	−	−	−	−	2.083±0.398
FAMILY	**1.00**	0.78	**1.00**	0.90	−	−	−	−	−	−	1.979±0.361
GHSM-2	**1.00**	0.86	**1.00**	**0.95**	−	−	−	−	−	−	2.081±0.336
AIL-3	**1.00**	0.82	**1.00**	0.85	0.92	0.88	−	−	−	−	2.087±0.280
GHSM-3	**1.00**	0.94	**1.00**	0.92	**0.95**	0.93	−	−	−	−	1.907±0.326
AIL-4	**1.00**	0.84	**1.00**	0.89	0.90	0.94	0.70	0.97	−	−	1.835±0.296
GHSM-4	**1.00**	0.88	**1.00**	0.89	0.88	0.92	**0.88**	**0.98**	−	−	1.813±0.369
AIL-5	**1.00**	0.61	**1.00**	0.68	0.92	0.84	0.58	0.96	0.25	**0.99**	2.074±0.327
GHSM-5	**1.00**	0.91	**1.00**	0.81	0.65	**0.96**	0.78	**0.98**	**0.80**	**0.99**	**1.793±0.377**

Table 3: **The statistic of the effects in the synthetic study. '#' indicates 'the number of'.**

	1st	2nd	3rd	4th	5th
# effects	20	190	1140	4845	15504
# non-zeros	10	5	6	4	2

5.1.2 Results and Analysis

The detailed results are shown in Table 2. The compared methods can be divided into 5 groups, i.e., {Lasso}, {AIL-2, w-hierNet, eWHL, s-hierNet, FAMILY, GHSM-2}, {AIL-3, GHSM-3}, {AIL-4, GHSM-4} and {AIL-5, GHSM-5}, according to the highest order of interactions that they can handle. From the results, we have several observations: (1) All the methods in comparison can correctly detect the useful covariates, since all the Sen. values in the column for the 1st-order interactions are 1's. Similar results are also observed for the 2nd-order interaction models; (2) In each synthetic dataset, the GHSM method has better recovery performance for different orders of interactions compared with other methods; (3) The prediction performance of each method is usually proportional to its capacity. That is, the methods that can learn higher-order interactions will achieve lower RMSE for prediction; (4) In all the three datasets, the GHSM method always has the best prediction performance and it significantly outperforms the Lasso and the existing second-order interaction models; (5) In the third synthetic dataset, the AIL-5 method performs even worse than the AIL-4 method. One possible reason is that the AIL method does not capture the sparse heredity structure, making it hardly detect the correct higher-order interactions. One evidence is that the AIL-5 method only recovers 58% and 25% of the correct 4th and 5th order interactions, respectively, while our GHSM-5 method achieves 78% and 80%.

The synthetic study demonstrates that as long as significant high-order interaction effects exist in the data, the high-order interaction methods will have better performance compared with the conventional second-order interaction methods, and the proposed GHSM method with high-order interactions can learn the sparse heredity structure and accurately detect those interactions, leading to improved prediction performance.

5.2 H3N2 Influenza A Virus Data

In this section, we study the application of the GHSM method on the antigenic sites identification problem.

5.2.1 Settings

Seasonal influenza A viruses pose great threats to public health, while the vaccination is the primary way to reduce this risk. An effective vaccination program requires an antigenic match between circulating viruses and vaccine strains to be used, and hence a timely identification of emerging influenza virus antigenic variants is critical to the success of influenza vaccination programs. Recent studies have suggested that multiple interactive antigenic sites mutations will significantly enhance the antigenic drift of the influenza viruses to new variants [17, 10, 23]. However, discovering the important interactive patterns among the antigenic sites is not trivial.

In this problem, each site is treated as a covariate of the antigenic distances which are the responses, hence identifying interactive patterns among multiple antigenic sites can be formulated as an interaction model. Here we apply the proposed GHSM method to identify interactive antigenic sites on an influenza H3N2 virus dataset [20, 23].[2] This dataset collects the results from the hemagglutination inhibition (HI) assays, which is a matrix recording the reaction values between 192 viruses in rows and a number of test serums in columns. The 192 H3N2 influenza A viruses are collected during year 2004 to 2007. These reaction values in the HI matrix describe the virus antigenicities, and the Euclidian difference between the reaction values of two viruses describes the antigenic distance between them. A large antigenic distance may induce an antigenic drift and sometimes cause influenza outbreaks due to viral escape from existing immunity. Therefore, accurately predicting the antigenic distances among viruses is a fundamental task. In this dataset, for each virus, its hemagglutinin (HA) protein sequence, i.e., a sequence of amino acid sites (covariates) that are responsible for antigenic changes, is also collected. The number of amino acid sites in the HA sequence of each virus is $d = 329$. By comparing the sites in any two HA sequences, we could obtain a difference vector, in which the unchanged positions have zero values and the mutated positions have integer values between 1 and 5, which is computed via the pattern-induced multi-sequence alignment (PIMA) scheme [20, 23]. Then, each pairwise difference vector is treated as a data sample in the data matrix $\mathbf{X}$, and each pairwise antigenic distance obtained from the HI reaction values is used as a response value in the target $\mathbf{y}$. So there are a total number of $\binom{192}{2} = 18336$ samples.

Table 4: Statistics of the influenza virus data.

	1st	2nd	3rd	4th
# effects	329	53,956	5,881,204	479,318,126
# valid effects	329	4368	37,857	157,462
rate	100%	8.1%	0.64%	0.03%
# samples	18,336			

Since there are 329 features, the number of interaction effects increases drastically with respect to the order. Table 4 gives statistics of the effects in this data. When we consider the 3rd- and 4th-order interactions, the dimension of the entire feature space is around hundreds of millions, which leads to heavy computational demand. Fortunately, in each data sample, we find that only few mutated positions have non-zero values and hence the data matrix $\mathbf{X}$ is sparse. As a consequence, a large number of the interaction effects are useless zero vectors since they are obtained via the product among sparse inputs and we can eliminate them before learning. The number of valid effects for each order is given in Table 4. Here we focus on up to the 4th-order interactions since higher-order interactions cannot bring too much performance improvement.

The entries in $\mathbf{X}$ and $\mathbf{Z}^{(k)}$'s are normalized into $[0, 1]$ and each entry in $\mathbf{y}$ is log-transformed and normalized into $[0, 1]$ as well. The regularization parameters for different methods are selected from a set $\{10^{-5}, 10^{-4}, \cdots, 10^{3}\}$ via the 5-fold cross validation. The parameter α in GHSM methods is chosen in the same way as that in the synthetic data. We randomly split the dataset into a training set and a test set by varying the training ratio from 10% to 90% at an interval of 20%. Each setting is repeated for 10 times. We report both the predictive RMSE on the test set and the running time for all the methods.

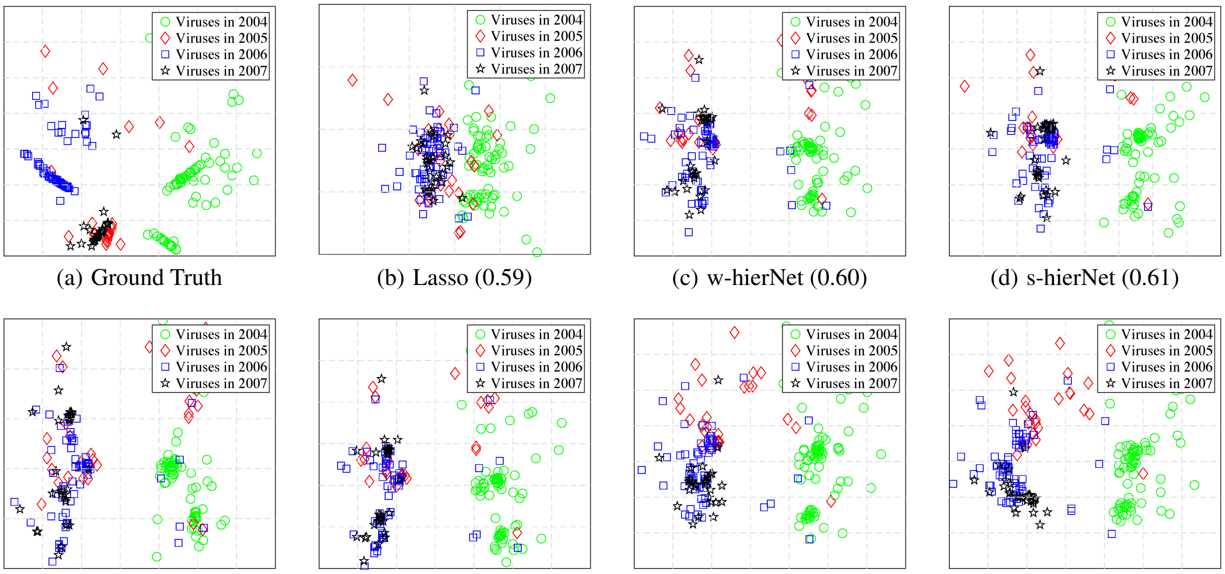
5.2.2 Results and Analysis

The predictive RMSE's for different methods are presented in Table 5. From the results, we observe that all the interaction models remarkably outperform the Lasso method, implying that incorporating the feature interactions is important in this dataset. Different from the synthetic case, the AIL methods do not show much improvement over the Lasso and this is possibly because the AIL methods can not make use of the sparse heredity structure and the interactions among the antigenic sites are much more complex than the synthetic case. For the second-order interaction models, the strong heredity based methods, i.e., the s-hierNet, FAMILY, and GHSM-2, show better performance than the weak heredity based methods including the w-hierNet and eWHL methods. This could be an evidence that the strong heredity structure is more useful for this problem. Similar to the results in the synthetic case, the GHSM method with higher-order interactions obtain more accurate prediction results and the GHSM-4 method performs the best in all settings.

In addition to the RMSE, we also provide visualizations for the prediction results of different methods via the antigenic cartography, which is an approach to visualize the virus antigenic evolution process on a 2D/3D space by using the antigenic distance among the viruses and has been widely used for virus antigenicity explanation since its first use by [18]. The idea in the antigenic cartography is to utilize the multi-dimensional scaling technique to obtain the coordinates of each virus on a 2D/3D space given the antigenic distance among them. We plot the embedding results in Fig. 2 when the training rate is 10% and use the predicted antigenic distance to reconstruct the cartography of all the viruses. In Fig. 2(a), the cartography using the true antigenic distance among the viruses is plotted in a 2D space, where each circle represents a virus and all the 192 viruses are divided into four different groups according to the year of their appearance. Figs. 2(b)-2(h) show the cartographies of different methods respectively. Since the cartographies of the AIL methods are similar to that of the Lasso and the eWHL and w-hierNet methods are two solutions of the same model, we omit their cartographies. By comparing with Fig. 2(a), we observe that the Lasso method can hardly reconstruct the antigenic distances among the viruses. The second-order interaction models including the w-hierNet, s-hierNet, FAMILY and GHSM-2 methods show clearer reconstruction but the viruses in different years are still hard to be distinguished. The group structure in the cartographies of the GHSM-3 and GHSM-4 methods is much better than others, which can be confirmed by the Pearson correlation coefficients reported in Fig. 2.

One advantage of the GHSM method is that we can identify important high-order interactions based on the model parameters. Specifically, we take the 4th-order interactions learned from the GHSM-4 method for example to see which 4th-order interactions are detected by the algorithm. We sort the magnitude of the coefficients for the 4th-order interactions in a descending order and then select the top 5 interactions, which are $\langle 157, 159, 242, 246\rangle$, $\langle 186, 193, 242, 246\rangle$, $\langle 94, 145, 189, 219\rangle$, $\langle 145, 189, 198, 219\rangle$ and $\langle 94, 145, 189, 198\rangle$. It is well-known in the H3N2 virus antigenicity analysis that there are 135 important antigenic sites out of the total 329 positions identified as the antibody binding sites, since these sites locate at the surface of the H3N2 virus protein structure and they are more likely to react with the sera. These antibody binding sites are further divided into 5 binding areas A-E according to their locations. Promisingly, we find that all of these detected positions in the 4th-order interactions belong to the binding areas. More interestingly, when we tag the binding areas for these sites in the selected interactions, we get the following patterns: ⟨B, B, D, D⟩, ⟨B, B, D, D⟩, ⟨E, A, B, D⟩, ⟨A, B, B, D⟩ and ⟨E, A, B, B⟩, respectively, from which we see that the antigenicity of the H3N2 virus is more likely to be controlled by the interactions among the sites in different binding regions simultaneously instead of in the same binding region. This observation is reasonable since multiple sites from different binding areas can accurately capture the 3D

[2] http://sysbio.cvm.msstate.edu/research/resources



(a) Ground Truth (b) Lasso (0.59) (c) w-hierNet (0.60) (d) s-hierNet (0.61)

(e) FAMILY (0.62) (f) GHSM-2 (0.64) (g) GHSM-3 (0.67) (h) GHSM-4 (**0.70**)

Figure 2: Cartographies of the prediction results of different methods when using 10% samples for training. The number in each bracket of figures (b)-(h) denotes the Pearson correlation coefficient between the true coordinates and the coordinates obtained from each method.

Table 5: The predictive RMSE's on the influenza virus data. '−' indicates that the corresponding algorithm does not return a result after running over 5 hours. The numbers in bold face denotes the best results.

Training	Covariates	Up to 2nd-order interactions					
Rate	Lasso	AIL-2	w-hierNet	eWHL	s-hierNet	FAMILY	GHSM-2
10%	0.3938±0.0076	0.3648±0.0150	0.2025±0.0011	0.2024±0.0023	0.1869±0.0040	0.1823±0.0029	0.1887±0.0018
30%	0.3914±0.0038	0.3622±0.0056	0.2018±0.0019	0.2021±0.0010	0.1855±0.0016	0.1818±0.0016	0.1884±0.0014
50%	0.3935±0.0020	0.3625±0.0027	0.2013±0.0013	0.2019±0.0009	0.1844±0.0040	−	0.1875±0.0014
70%	0.3927±0.0011	0.3634±0.0022	0.1989±0.0024	0.2015±0.0013	0.1835±0.0024	−	0.1873±0.0022
90%	0.3926±0.0027	0.3629±0.0024	0.1958±0.0011	0.2013±0.0022	0.1815±0.0018	−	0.1868±0.0022

Training	Up to 3rd-order interactions		Up to 4th-order interactions	
Rate	AIL-3	GHSM-3	AIL-4	GHSM-4
10%	0.3655±0.0143	0.1742±0.0014	0.3658±0.0163	**0.1737±0.0013**
30%	0.3638±0.0066	0.1735±0.0007	0.3642±0.0065	**0.1725±0.0013**
50%	0.3641±0.0031	0.1733±0.0015	0.3635±0.0034	**0.1714±0.0011**
70%	0.3640±0.0022	0.1732±0.0013	0.3636±0.0025	**0.1718±0.0015**
90%	0.3620±0.0022	0.1723±0.0031	0.3631±0.0029	**0.1697±0.0039**

structure (shape) of the virus. Fig. 3 shows the 3D structure of these binding areas with the sites from the selected interactions. So far, all of the above analysis is considerably useful in influenza vaccine strain selection, and it can significantly reduce the human labor efforts for serological characterization and will increase the probability of correct influenza vaccine candidate selection.

Moreover, we compare the training time of different HSM models on the entire influenza virus dataset and report the results in Table 6. For the 2nd-order HSM, the strong heredity based methods, i.e., the s-hierNet and FAMILY, are computationally much more expensive than the weak heredity based methods such as the w-hierNet and eWHL. The FAMILY method fails to give the solution in reasonable time. Actually, the FAMILY method is computationally intractable even when using only 50% of the samples for training (refer to Table 5). The eWHL method is very efficient since it is specifically designed for the w-hierNet model. The proposed GHSM-2 method, a strong heredity based method, is much more efficient than the s-hierNet and FAMILY methods and comparable with the eWHL algorithm. By increasing the order of interactions, the GHSM-3 and GHSM-4 methods are still very efficient compared with the w-hierNet, s-hierNet and FAMILY methods, while it has much better performance than those methods.

6. CONCLUSION

In this paper, we proposed a generalized hierarchical sparse model to learn arbitrary-order interactions contained in the data via the proposed arbitrary-order heredity structure. An efficient algorithm was developed by decoupling the variables in the complex constraint and all the subproblems have efficient analytical solutions. Empirical results show the effectiveness of the proposed method.

When considering high-order interactions, if the data matrix are not sparse like the influenza virus data, the number of high-order interactions still increases exponentially with respect to the order and solving the GHSM will become intractable even for small d and K. One possible direction to solve this problem is to conduct dimensionality reduction methods before learning the GHSM model via, for example, the feature screening technique. We are also interested in applying the GHSM methods to other biological problems, such as the cancer microarray analysis, to detect important interactions.

Acknowledgments

This work was partially supported by NSF IIS-1250985, NSF IIS-1407939, NIH R01AI116744, NSFC-61305071, NSFC-61473087 and NSF of Jiangsu Province (BK20141340).

Table 6: **Training time of all HSMs on the entire flu virus data. '−' indicates that the algorithm does not return a result after running over 5 hours.**

	2nd-order interaction models					Higher-order interaction models	
	w-hierNet	eWHL	s-hierNet	FAMILY	GHSM-2	GHSM-3	GHSM-4
Training time (in seconds)	205.6	11.3	5876.9	−	21.3	34.3	76.4

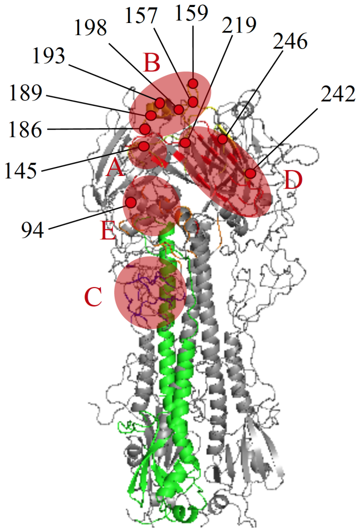


Figure 3: **The 3D structure of the H3N2 virus. The red regions denote the anti-body binding regions A-E. The sites from the selected interactions are also labeled.**

7. REFERENCES

[1] J. Bien, J. Taylor, and R. Tibshirani. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

[2] R. J. Cadoret, W. R. Yates, G. Woodworth, and M. A. Stewart. Genetic-environmental interaction in the genesis of aggressivity and conduct disorders. *Archives of General Psychiatry*, 52(11):916–924, 1995.

[3] N. H. Choi, W. Li, and J. Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

[4] J. F. Dawson and A. W. Richter. Probing three-way interactions in moderated multiple regression: development and application of a slope difference test. *Journal of Applied Psychology*, 91(4):917, 2006.

[5] T. C. Eley, K. Sugden, A. Corsico, A. M. Gregory, P. Sham, P. McGuffin, R. Plomin, and I. W. Craig. Gene–environment interaction analysis of serotonin system markers with adolescent depression. *Molecular Psychiatry*, 9(10):908–915, 2004.

[6] J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, Berlin, 2001.

[7] P. Gong, C. Zhang, Z. Lu, J. Z. Huang, and J. Ye. A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. In *ICML*, 2013.

[8] L. Han and Y. Zhang. Learning tree structure in multi-task learning. In *KDD*, pages 397–406, 2015.

[9] A. Haris, D. Witten, and N. Simon. Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*, 2014.

[10] J.-W. Huang, C.-C. King, and J.-M. Yang. Co-evolution positions and rules for antigenic variants of human influenza a/h3n2 viruses. *BMC Bioinformatics*, 10(Suppl 1):S41, 2009.

[11] M. Lim and T. Hastie. Learning interactions through hierarchical group-lasso regularization. *arXiv preprint arXiv:1308.2719*, 2013.

[12] Y. Liu, J. Wang, and J. Ye. An efficient algorithm for weak hierarchical lasso. In *KDD*, pages 283–292, 2014.

[13] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*, volume 821. John Wiley & Sons, 2012.

[14] D. D. Pollock, W. R. Taylor, and N. Goldman. Coevolving protein residues: maximum likelihood identification and relationship to structure. *Journal of Molecular Biology*, 287(1):187–198, 1999.

[15] P. Radchenko and G. M. James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105(492):1541–1553, 2010.

[16] Y. She and H. Jiang. Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*, 2014.

[17] A. C.-C. Shih, T.-C. Hsiao, M.-S. Ho, and W.-H. Li. Simultaneous amino acid substitutions at antigenic sites drive influenza a hemagglutinin evolution. *PNAS*, 104(15):6283–6288, 2007.

[18] D. J. Smith, A. S. Lapedes, J. C. de Jong, T. M. Bestebroer, G. F. Rimmelzwaan, A. D. Osterhaus, and R. A. Fouchier. Mapping the antigenic and genetic evolution of influenza virus. *Science*, 305(5682):371–376, 2004.

[19] M. J. Somers. Organizational commitment, turnover and absenteeism: An examination of direct and interaction effects. *Journal of Organizational Behavior*, 16(1):49–58, 1995.

[20] H. Sun, J. Yang, T. Zhang, L.-P. Long, K. Jia, G. Yang, R. J. Webby, and X.-F. Wan. Using sequence data to infer the antigenicity of influenza virus. *MBio*, 4(4):e00230–13, 2013.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[22] X. Yan and J. Bien. Hierarchical sparse modeling: a choice of two regularizers. *arXiv preprint arXiv:1512.01631*, 2015.

[23] J. Yang, T. Zhang, and X.-F. Wan. Sequence-based antigenic change prediction by a sparse learning method incorporating co-evolutionary information. *PLOS One*, 9(9):e106660, 2014.

[24] M. Yuan, V. R. Joseph, and H. Zou. Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757, 2009.

[25] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

[26] P. Zhao, G. Rocha, and B. Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.