Knowledge Based Factorized High Order Sparse Learning Models

Sanjay Purushotham [*] Martin Renqiang Min [†] C.-C. Jay Kuo [‡] Mark Gerstein [§]

Abstract

We propose a scalable knowledge based high order sparse learning framework termed as **Group F**actorized **H**igh order **I**nteractions **M**odel (Group FHIM) for identifying discriminative feature groups and high-order feature group interactions in classification problems. Our factorization technique allows us to incorporate the domain knowledge such as grouping of features directly into the decomposition factors. Unlike previous sparse learning approaches, our model can recover both the discriminative feature groups and the pairwise feature group interactions accurately without enforcing any hierarchical feature constraints. We show that our Group FHIM estimator is asymptotically optimal. Experiments on synthetic and real datasets show that our model outperforms the state-of-the-art sparse learning techniques, and it provides 'interpretable' high-order feature group interactions for gene expression prediction and peptide-MHC I binding prediction.

1 Introduction

In machine learning and data mining, reliably identifying interpretable discriminative interactions among high-dimensional input features with limited training data remains an unsolved problem. For example, a major challenge in biomarker discovery and personalized medicine is to identify gene/protein interactions and their relations with other physical factors in medical records to predict the health status of patients. However, we often have limited patient samples but hundreds of millions of feature interactions to consider. Recently, some researchers tried to solve this problem by making sparsity and hierarchical constraint assumptions to find discriminative features and their interactions. Hierarchical constraints for high-order feature interactions are suitable for some real-world problems but are too stringent for some others. In real-world applications, we often have abundant prior information about input features that can be readily obtained from a lot of knowledge bases, especially in this big data era. To address the above challenging problem of identifying high order feature interactions, we need to build scalable models by incorporating existing knowledge about input features into the model construction process.

In this paper, we propose a novel knowledge-based sparse learning framework based on weight matrix factorizations and ℓ_1/ℓ_2 regularization for identifying discriminative high-order feature group interactions in logistic regression and large-margin models, and we study theoretical properties for the same. Experimental results on synthetic and real-world datasets show that our method outperforms the state-of-the-art sparse learning techniques, and it provides 'interpretable' blockwise high-order feature interactions for gene expression prediction and peptide-MHC I protein binding prediction. Our proposed sparse learning framework is quite general, and can be used to identify any discriminative complex system input interactions that are predictive of system outputs given limited high-dimensional training data.

Our contributions are as follows: (1) We propose a method capable of simultaneously identifying both informative discriminative feature groups and discriminative high order feature group interactions in a sparse learning framework by incorporating domain knowledge; (2) Our method works on high-dimensional input feature spaces with much more features than data samples, which is typical for biomedical applications, (3) Our method has interesting theoretical properties for generalized linear regression models; (4) The feature group interactions identified by our method leads to better understanding of peptide-MHC I protein interaction and gene transcriptional regulation.

2 Related Work

Feature selection is a classical problem and has been well studied using Kernel methods such as Support Vector Machines (SVMs) [18], Multiple Kernel Learning [11], Gaussian Processes [5] and Regularization methods such as Lasso [19], Group Lasso [21] etc. Even though there has been extensive research in these feature selection techniques, they have mainly focused on identifying individual discriminative features. Kernel methods have been used to model high order feature in-

[*] University of Southern California, *Email: spurusho@usc.edu*
[†] NEC Labs of America, *Email: renqiang@nec-labs.com*
[‡] University of Southern California, *Email: cckuo@sipi.usc.edu*
[§] Yale University, *Email: mark.gerstein@yale.edu*

teractions, but they only help to identify which orders are important rather than finding the relevant high order feature interactions. Recently, regularization methods have become very popular for feature selection because they are suited for the high dimensional problems. Many regularization methods focus on identifying discriminative features or groups of discriminative features based on ℓ_1 penalty, Group penalty, Trace-norm [6] penalty, (k, q) penalty [17] or Dirty model [8]. More recent approaches [3], [1], [14] are aimed at recovering not only the discriminative features but also high order feature interactions in regression models by enforcing strong and/or weak heredity (hierarchical) constraints. In strong heredity, a feature interaction term is included in the model only if the corresponding features are also included in the model, while in weak heredity, a feature interaction term is included when one of the features is included in the model [1]. Even though hierarchical constraints help model interpretability in some applications, recent studies in bioinformatics and related areas have shown that feature interactions need not follow heredity constraints for manifestation of the diseases; and thus the above approaches based on heredity constraints have limited chance of recovering relevant interactions in these areas.[15] proposed an efficient way to identify combinatorial interactions among interactive genes in complex diseases by using prior information such as gene ontology. However, they also make hereditary assumptions which limits their model's capacity at capturing all the important high order interactions. Thus, all these previous approaches are very unlikely to recover 'interpretable' blockwise high order feature and feature group interactions for prediction due to heredity constraints or they do not incorporate the existing domain knowledge. This motivates us to develop new efficient knowledge based techniques to capture the important 'blockwise' high-order feature and feature group interactions without making heredity assumptions.

Recently, in our previous work [16] we proposed **F**actoriz-ed **H**igh order **I**nteractions **M**odel (**FHIM**) to identify high order feature interactions in regression models in a greedy way based on ℓ_1 penalty on features and without assuming heredity constraints. This paper generalizes the sparse learning framework introduced in [16] with the following new and significant contributions: 1) In this paper, we show how to incorporate domain knowledge into the sparse learning framework using knowledge-based factorization technique and regularization penalties, 2) We show state-of-the-art results on 3 real world datasets to showcase the advantage of capturing prior information in our sparse learning framework, 3) We show that our Group FHIM still has the nice theoretical properties as FHIM.

The remainder of the paper is organized as follows: in section 3 we discuss our problem formulation and relevant notations used in the paper. In section 4, we discuss the main idea of this paper and present the optimization method used in our sparse learning framework. We give an overview of theoretical properties in section 5. In section 6, we discuss our experimental setup and present our results on synthetic and real datasets. Finally, in section 7 we conclude the paper with discussion and future research directions.

3 Notations and Problem Formulation

For any vector $\boldsymbol{w}$, let $||\boldsymbol{w}||_2$ denote the Euclidean norm of $\boldsymbol{w}$, and $supp(\boldsymbol{w}) \subset [1, p]$, denote the support of $\boldsymbol{w}$, i.e. the set of features $i \in [1, p]$ with $w_i \neq 0$. A group of features is a subset $g \subset [1, p]$. The set of all possible groups is the power set of $[1, p]$ and let us donate it as $\mathcal{P}$. Let $\mathcal{G} \subset \mathcal{P}$ denote a set of groups of features. In our paper, the domain knowledge is presented in terms of $\mathcal{G}$. For any vector $\boldsymbol{w} \in \mathbb{R}^p$, and any group $g \in \mathcal{G}$, let $\boldsymbol{w}_g$ denote a vector whose entries are the same as $\boldsymbol{w}$ for the features in g and 0 for other features. Let $\boldsymbol{W}_g$ denote a matrix of size $p \times p$ for some $g \in \mathcal{G}$ and the entries of $\boldsymbol{W}_g$ are non-zero for corresponding column entries in g (i.e. $\boldsymbol{W}_g^{ij} \neq 0$ for $g \in \mathcal{G}$ and 0 otherwise). Let $\mathcal{V}_\mathcal{G} \in \mathbb{R}^{p \times \mathcal{G}}$ denote a set of $N_\mathcal{G}$ tuples of vector $\boldsymbol{v} = (v_g)_{g \in \mathcal{G}}$, where each v_g is a separate vector in $\mathbb{R}^p$, with $supp(v_g) \subset g, \forall g \in \mathcal{G}$. If two groups overlap then they share at least one feature in common.

Let $\{(\mathbf{X}^{(i)}, y^{(i)})\}, i \in [1, n]$ represent a training set of n samples and p features (predictors), where $\mathbf{X}^{(i)} \in \mathbb{R}^p$ is the i^{th} instance (column) of the design matrix $\mathbf{X}$ and $y(i) \in \{-1, 1\}$ is the i^{th} instance of response variable (output) $\mathbf{y}$. Let $\{\boldsymbol{\beta}, \boldsymbol{\beta}_g\} \in \mathbb{R}^p$ be the weight vector associated with single features (also called main effects) and feature groups respectively, and $\beta_0 \in \mathbb{R}$ be the bias term. Note, $\boldsymbol{\beta} = \sum_{g \in \mathcal{G}} \boldsymbol{\beta}_g$. Let $\mathbf{W}$ be the weight matrix associated with the pairwise feature group interactions and let $\mathbf{W}_{OD}$ be the weight matrix associated with only the pairwise feature group interactions without self interactions. $\mathbf{W}_{OD}$ is an off-diagonal matrix and is given by equation (4.7).

In this paper, we study the problem of identifying the discriminative feature groups $\boldsymbol{\beta}_g$ and the pairwise feature group interactions $\mathbf{W}_{OD}$ in classification settings, when domain knowledge such as grouping of features ($\mathcal{G}$) is given, and without making any heredity assumptions. For the classification settings, we can model the output in terms of features and their high order interactions using logistic regression model or large-margin models. Here we consider both these popular classifiers. A logistic regression model with pairwise interactions can be written as follows.

$$p(y^{(i)}|\mathbf{X}^{(i)}) = \frac{1}{1+\exp(-y^{(i)}(\boldsymbol{\beta}^T\mathbf{X}^{(i)}+\mathbf{X}^{(i)T}\mathbf{W}_{OD}\mathbf{X}^{(i)}+\beta_0))} \tag{3.1}$$

The corresponding loss function (the sum of the negative log-likelihood of the training data) is given by

$$\begin{aligned} L_{LogReg}(\boldsymbol{\beta},\mathbf{W}_{OD},\beta_0) = \sum_{i=1}^{n} &\log(1+\exp(-y^{(i)}(\boldsymbol{\beta}^T\mathbf{X}^{(i)} \\ &+\mathbf{X}^{(i)T}\mathbf{W}_{OD}\mathbf{X}^{(i)}+\beta_0)). \end{aligned} \tag{3.2}$$

Similarly, we can solve the classification problem with high order interactions using large margin formulation with Hinge Loss as follows

$$\begin{aligned} L_{Hinge}(\boldsymbol{\beta},\mathbf{W}_{OD},\beta_0) = \sum_{i=1}^{n} &max(0, 1-(y^{(i)}(\boldsymbol{\beta}^T\mathbf{X}^{(i)} \\ &+\mathbf{X}^{(i)T}\mathbf{W}_{OD}\mathbf{X}^{(i)}+\beta_0)). \end{aligned} \tag{3.3}$$

4 Group FHIM

Here, we present our optimization-driven knowledge based sparse learning framework to identify discriminative feature groups and pairwise feature-group interactions (blockwise interactions) for the classification problems of previous section. For simplicity, here we consider that the groups do not overlap. A natural way to recover the feature groups and their interactions is by regularization as shown below.

$$\begin{aligned} \{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{W}}\} = \arg\min_{\boldsymbol{\beta},\boldsymbol{W}} \mathcal{L}(\boldsymbol{\beta},\mathbf{W}) + \lambda_\beta \sum_{g\in\mathcal{G}} ||\boldsymbol{\beta}_g||_2 \\ + \lambda_W \sum_{g\in\mathcal{G}} ||vec(\boldsymbol{W}_g)||_2 \end{aligned} \tag{4.4}$$

where $vec(\boldsymbol{W}_g)$ is the vectorization of the group block matrix $\boldsymbol{W}_g$. When the number of input features is huge (e.g. biomedical applications), it is practically impossible to explicitly consider pairwise or even higher-order interactions among all the input feature groups based on simple ℓ_1-penalty or Group Lasso penalty. To solve this problem, we propose a novel way to factorize the block-wise interaction weight matrix $\mathbf{W}$ as sum of K rank-one matrices. Each rank-one matrix is represented by an outer product of two identical vectors (termed as rank-one factors) with the grouping structure imposed on these vectors. The feature group interactions of $\mathbf{W}$ can be effectively captured by the grouping on the rank-one factors. A feasible decomposition of blockwise $\mathbf{W}$ is shown below

$$\mathbf{W} = \sum_{k=1}^{K}(\sum_{g\in\mathcal{G}} \boldsymbol{a}_{kg}) \otimes (\sum_{g\in\mathcal{G}} \boldsymbol{a}_{kg})$$

where $\otimes$ represents the tensor product/outer product and $\boldsymbol{a}_k$ is a rank-one factor of $\mathbf{W}$ and is given by $\boldsymbol{a}_k = \sum_{g\in\mathcal{G}} \boldsymbol{a}_{kg}$. The above decomposition is feasible since each rank-one matrix decomposition of $\mathbf{W}$ can be represented as weighted combinations of the group block matrices $\mathbf{W}_g$.

Now, we can rewrite the optimization problem (4.4) to identify the discriminative feature groups and pairwise feature group interactions by using the grouped rank-one factors as follows,

$$\{\hat{\boldsymbol{\beta}},\hat{\boldsymbol{a}}_k\} = \arg\min_{\boldsymbol{a}_k,\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta},\mathbf{W}_{OD}) + \mathcal{P}_\lambda(\boldsymbol{\beta},\boldsymbol{a}_k) \tag{4.5}$$

where,

$$\mathcal{P}_\lambda(\boldsymbol{\beta},\boldsymbol{a}_k) = \lambda_\beta \sum_{g\in\mathcal{G}} ||\boldsymbol{\beta}_g||_2 + \sum_k \lambda_{a_k} \sum_{g\in\mathcal{G}} ||\boldsymbol{a}_{kg}||_2 \tag{4.6}$$

and

$$\begin{aligned} \mathbf{W}_{OD} = \sum_{k=1}^{K}(\sum_{g\in\mathcal{G}} \boldsymbol{a}_{kg}) \otimes (\sum_{g\in\mathcal{G}} \boldsymbol{a}_{kg}) \\ - \mathcal{D}(\sum_{k=1}^{K}(\tilde{a}_{k,i}^2)_{i\in[1,p]}) \end{aligned} \tag{4.7}$$

where $\hat{\boldsymbol{\beta}},\hat{\boldsymbol{a}}_k$ represent the estimated parameters of our model, $\mathcal{D}$ is a diagonalizing matrix operator which returns a $p \times p$ diagonal matrix, and $\tilde{a}_{k,i}$ is the i^{th} component of $\boldsymbol{a}_k$.

Let Q represent the objective function (loss function with the regularization penalties) i.e. the right hand side of the equation (4.5). We replace $\mathcal{L}$ in (4.5) by $L_{LogReg}(\boldsymbol{\beta},\mathbf{W}_{OD},\beta_0)$ for logistic regression, and by $L_{Hinge}(\boldsymbol{\beta},\mathbf{W}_{OD},\beta_0)$ for large-margin classification. We call our model **Group Factorization based High-order Interaction Model (Group FHIM)**. In section 4.1 we present a greedy alternating optimization algorithm to solve our optimization problem. Note that we use $\mathbf{W}_{OD}$ in equation (4.5) instead of $\mathbf{W}$. Although the original $\mathbf{W}$ is a sum of K rank-one matrix with the maximum rank K, the actual rank of $\mathbf{W}_{OD}$ is often much larger than K. However, $\mathbf{W}$ and the off-diagonal $\mathbf{W}_{OD}$ define the same interaction block-wise patterns between different input features. In practice, we often focus on identifying interpretable discriminative high-order interactions between different features instead of uninteresting self-interactions. Moreover, removing diagonal elements of $\mathbf{W}$ has the advantage of eliminating the interference between optimizing $\boldsymbol{\beta}$ and optimizing $\boldsymbol{a}_k$'s for binary input feature vectors, which greatly helps our alternating optimization procedure and often results in much better local optimum in practice. Our empirical studies also show that, even for continuous input features, $\mathbf{W}_{OD}$ often result in faster parameter learning and better local optima. Therefore, we used $\mathbf{W}_{OD}$ instead of $\mathbf{W}$ in the objective functions of both FHIM and Group FHIM for all the experiments in this paper.

Remark: *Overlapping Group FHIM* - The non overlapping group structure used in Group FHIM limits its applicability in practice. Hence, we propose an extension of Group FHIM to overlapping groups case and call our method Overlapping Group FHIM (denoted by OvGroup FHIM). In OvGroup FHIM, we consider the overlapping group penalty [7] instead of the ℓ_1/ℓ_2 penalty used in Group FHIM. The overlapping group penalty for $\boldsymbol{a}_k$ is given below.

$$\Omega^{\mathcal{G}}_{overlap}(\boldsymbol{a}_k) = \inf_{v \in \mathcal{V}_{\mathcal{G}}, \sum_{g \in \mathcal{G}} v_g = a_k} \sum_{g \in \mathcal{G}} ||v_g|| \tag{4.8}$$

4.1 Greedy Alternating Optimization

The optimization problem in Equation 4.5 is convex in $\boldsymbol{\beta}$ but non-convex in $\boldsymbol{a}_k$. The non-convexity property of our optimization problem makes it is difficult to propose an optimization strategy which guarantees convergence to global optima. Here, we propose a greedy alternating optimization approach (Algorithm 1) to find a local optima for our problem. We use the Spectral Projected Gradient method for solving our optimization problems (Line 4 and 5) since we found through experiments that it is much faster than other popular approaches such as Quasi-Newton methods.

Algorithm 1 Greedy Alternating Optimization

```
1: Initialize $\boldsymbol{\beta}$ to $\boldsymbol{\beta}_{LASSO}$, $K = 1$ and $\boldsymbol{a}_K = \mathbf{1}$
2: While (K==1) $OR$ ($\boldsymbol{a}_{K-1} \neq \mathbf{0}$ $for$ $K > 1$)
3:     Repeat until convergence
4:         $a^t_{K,j} = \arg\min_j Q((a^t_{K,1}, ..., a^t_{K,j-1}, a^{t-1}_{K,j+1}, ... a^{t-1}_{K,p}), \boldsymbol{\beta}^{t-1})$
5:         $\beta^t_j = \arg\min_j Q(\beta^t_1, ..., \beta^t_{j-1}, \beta^{t-1}_{j+1}, \beta^{t-1}_p), \boldsymbol{a_K}^t)$
6:     End Repeat
7:     K = K + 1; $\boldsymbol{a}_K = \mathbf{1}$
8: End While
9: Return $\boldsymbol{a}_K$ and $\boldsymbol{\beta}$ which has the least loss function.
```

5 Theoretical Properties

In this section, we study the asymptotic behavior of our proposed Group FHIM for the likelihood based generalized linear regression models (eg. logistic regression model). The theorems shown here are similar to the ones in our previous work [16]. However, in this paper, we show that the asymptotic properties still holds even with the regularization penalty on rank-one factors used in our Group FHIM estimator.

Problem Setup: Assume that the data $\mathbf{V}_i = (\mathbf{X}_i, y_i), i = 1, ...n$ are collected independently and Y_i has a density of $f(Z(\mathbf{X}_i), y_i)$ conditioned on $\mathbf{X}_i$, where Z is a known regression function with grouped main effects and all possible pairwise group interactions. Let $\boldsymbol{\beta}^*_h$ and $\boldsymbol{a}^*_{k,h}$ denote the underlying true parameters satisfying block-wise properties implied by our factorization. Let $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{\alpha}^{*T})^T$, where $\boldsymbol{\beta}^* = (\beta^*_h), \boldsymbol{\alpha}^* = (a^*_{k,h}), k = 1, ..., K; h = 1, ..., |\mathcal{G}|$ (Note: $\boldsymbol{\theta}^*$ is $p(K+1) \times 1$). We consider the estimates for Group FHIM as $\hat{\boldsymbol{\theta}}_n$:

$$\begin{aligned} \hat{\boldsymbol{\theta}}_n &= \arg\min_{\theta} Q_n(\boldsymbol{\theta}) \\ &= \arg\min_{\theta} -\frac{1}{n} \sum_{i=1}^{n} (L(Z(\mathbf{X}_i), y_i) + \lambda_\beta \sum_h ||\boldsymbol{\beta}_h||_2 \\ &\quad + \sum_k \lambda_{\alpha_k} \sum_h ||\boldsymbol{\alpha_{k,h}}||_2 \end{aligned} \tag{5.9}$$

where $L(Z(\mathbf{X}_i), y_i)$ is the loss function of generalized linear regression models with pairwise feature group interactions. In the case of logistic regression, $Z(\cdot)$ takes the form of Equation (3.1) and $L(\cdot)$ takes the form of Equation (3.2). Now, let us define

$$\begin{aligned} \mathcal{A}_1 &= \{h : \beta^*_h \neq 0\} \\ \mathcal{A}_2 &= \{(k, h') : \alpha^*_{k,h'} \neq 0\}, \\ \mathcal{A} &= \mathcal{A}_1 \cup \mathcal{A}_2 \end{aligned} \tag{5.10}$$

where $\mathcal{A}_1$ contains the indices of the groups of main terms which correspond to the non-zero true group coefficients, and similarly $\mathcal{A}_2$ contains the indices of the factorized group interaction terms whose true group coefficients are non-zero. Let us define

$$\begin{aligned} a_n &= \max\{\lambda^h_\beta, \lambda^{h'}_{\alpha_k} : h \in \mathcal{A}_1, (k, h') \in \mathcal{A}_2\} \\ b_n &= \min\{\lambda^h_\beta, \lambda^{h'}_{\alpha_k} : h \in \mathcal{A}^c_1, (k, h') \in \mathcal{A}^c_2\} \end{aligned} \tag{5.11}$$

where $\mathcal{A}^c_1$ is the complement of set $\mathcal{A}_1$. Now, we show that our model possesses the oracle properties for $n \to \infty$ with fixed p under some regularity conditions. Note, the asymptotic properties for $p_n \to \infty$ as $n \to \infty$ will be addressed in our future work.

5.1 Asymptotic Oracle Properties when $n \to \infty$

The asymptotic properties when sample size increases and the number of predictors is fixed are described in the following theorems. We will show that Group FHIM possesses oracle properties under certain regularity conditions (R1)-(R3) shown below. Let Ω denote the parameter space for $\boldsymbol{\theta}$.

(R1) The observations $\mathbf{V}_i : i = 1, ..., n$ are independent and identically distributed with a probability density $f(\mathbf{V}, \boldsymbol{\theta})$, which has a common support. We assume the density f satisfies the following equations:

$$E_{\boldsymbol{\theta}}\Big[\frac{\partial \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \theta_j}\Big] = 0 \qquad \text{for } j = 1, ..., p(K+1),$$

and

$$\begin{aligned} \mathbf{I}_{jk}(\boldsymbol{\theta}) =& E_{\boldsymbol{\theta}}\Big[\frac{\partial \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \theta_k}\Big] \\ =& E_{\boldsymbol{\theta}}\Big[-\frac{\partial^2 \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \theta_j \partial \theta_k}\Big] \end{aligned}$$

(R2) The Fisher Information Matrix

$$\mathbf{I}(\boldsymbol{\theta}) = E\Big[\Big(\frac{\partial \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big)\Big(\frac{\partial \log f(\mathbf{V}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big)^T\Big]$$

is finite and positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

(R3) There exists an open set ω of Ω that contains the true parameter point $\boldsymbol{\theta}^*$ such that for almost all $\mathbf{V}$ the density $f(\mathbf{V}, \boldsymbol{\theta})$ admits all third derivatives $(\partial^3 f(\mathbf{V}, \boldsymbol{\theta}))/(\partial\theta_j\ \partial\theta_k \partial\theta_l)$ for all $\boldsymbol{\theta} \in \omega$ and any $j, k, l = 1, ..., p(K+1)$. Furthermore, there exist functions M_{jkl} such that

$$\left|\frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \log f(\mathbf{V}, \boldsymbol{\theta})\right| \leq M_{jkl}(\mathbf{V}) \qquad \text{for all } \boldsymbol{\theta} \in \omega$$

where $m_{jkl} = E_{\boldsymbol{\theta}^*}[M_{jkl}(\mathbf{V})] < \infty$. These regularity conditions are the existence of common support and first, second derivatives for $f(\mathbf{V}, \boldsymbol{\theta})$; Fisher Information matrix being finite and positive definite; and existence of bounded third derivative for $f(\mathbf{V}, \boldsymbol{\theta})$. These regularity conditions guarantee asymptotic normality of the ordinary maximum likelihood estimates [12].

Theorem 5.1. *Assume $a_n = o(1)$ as $n \to \infty$. Then under regularity conditions (R1)-(R3), there exists a local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta^*}|| = O_P(n^{-1/2} + a_n)$*

Remark. Theorem 5.1 implies that when the tuning parameters associated with the non-zero coefficients of grouped main effects and grouped pairwise interactions tend to 0 at a rate faster than $n^{-1/2}$, then there exists a local minimizer of $Q_n(\boldsymbol{\theta})$, which is $\sqrt{n}-$consistent (the sampling error is $O_p(n^{-1/2})$).

Theorem 5.2. *Assume $\sqrt{n}a_n \to 0, \sqrt{n}b_n \to \infty$ and $P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0) \to 1$. Then under the regularity conditions (R1)-(R3), the component $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{A}}}$ of the local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ (given in theorem 5.1) satisfies*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}^*_{\mathcal{A}}) \to_d N(0, \mathbf{I}^{-1}(\boldsymbol{\theta}^*_{\mathcal{A}})),$$

where $\mathbf{I}(\boldsymbol{\theta}^*_{\mathcal{A}})$ is the Fisher information matrix of $\boldsymbol{\theta}_{\mathcal{A}}$ at $\boldsymbol{\theta}_{\mathcal{A}} = \boldsymbol{\theta}^*_{\mathcal{A}}$ assuming that $\boldsymbol{\theta}^*_{\mathcal{A}^c} = 0$ is known in advance.

Remark. Theorem 5.2 shows that our model estimates the non-zero coefficients of the true model with the same asymptotic distribution as if the zero coefficients were known in advance. Based on theorems 5.1 and 5.2, we can say that our group FHIM estimator has the oracle property, i.e. it is asymptotically optimal, namely unbiased and efficient, when the tuning parameters satisfy the conditions $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$. To satisfy these conditions, we have to consider adaptive weights $w_j^{\beta}, w_l^{\alpha_k}$ [23] for our tuning parameters $\lambda_\beta, \lambda_{\alpha_k}$. Thus, our tuning parameters are:

$$\lambda_j^{\beta} = \frac{\log n}{n}\lambda_\beta w_j^{\beta}, \qquad \lambda_l^{\alpha_k} = \frac{\log n}{n}\lambda_{\alpha_k} w_l^{\alpha_k}$$

Note. Please see Supplementary materials for proofs.

5.2 Properties of Overlapping Group FHIM

Lemma 5.1. *$\beta \mapsto \Omega^{\mathcal{G}}_{overlap}(\beta)$ is a norm.*

Proof. : Lemma 1 [7]

Overlapping Group FHIM (OvGroup FHIM) can be realized using a non-overlapped Group FHIM. Let us form $\tilde{\mathbf{X}} \in \mathbb{R}^{n \times \sum |g|}$ by the concatenation of copies of the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ restricted to a certain group g, i.e. $\tilde{\mathbf{X}} = [\mathbf{X}_{g1}, .., \mathbf{X}_{g|\mathcal{G}|}]$ and $\mathcal{G} = g_1, .., g_{\mathcal{G}}; \tilde{\mathbf{v}} = (\tilde{v}^T_{g1}, .., \tilde{v}^T_{g|\mathcal{G}|})$, i.e. $\tilde{\mathbf{v}} \in \mathbb{R}^{\sum |g|}$ and with $\tilde{v}_g = (v_{gi})_{i \in g}$. Let the empirical risk for OvGroup FHIM and equivalent Group FHIM be represented by $R(.)$ and $\tilde{R}(.)$ respectively. Therefore, $R(\boldsymbol{a}_k) = \tilde{R}(\boldsymbol{X}^T\boldsymbol{a}_k\boldsymbol{a}_k^T\boldsymbol{X})$ and $R(\boldsymbol{\beta}) = \tilde{R}(\boldsymbol{X}\beta)$ respectively.

Theorem 5.3. *(i) $R(\boldsymbol{a}_k) = \tilde{R}(\tilde{\boldsymbol{X}}^T\tilde{v_g}\tilde{v_g}^T\tilde{\boldsymbol{X}})$ and (ii) $R(\boldsymbol{\beta}) = \tilde{R}(\tilde{\boldsymbol{X}}\tilde{v_g})$*

Proof. : We prove (i) here. Proof for (ii) is similar.

$$\begin{aligned}
R(\boldsymbol{a}_k) =& \tilde{R}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})\\
=& \tilde{R}(\boldsymbol{X}^T\boldsymbol{a}_k\boldsymbol{a}_k^T\boldsymbol{X})\\
=& \tilde{R}(\boldsymbol{X}^T(\sum_g a_{kg})(\sum_g a_{kg})^T\boldsymbol{X})\\
=& \tilde{R}(\tilde{\boldsymbol{X}}^T\tilde{v_g}\tilde{v_g}^T\tilde{\boldsymbol{X}})
\end{aligned}$$

Remark. Theorem 5.3 shows that empirical risk minimization of Overlapping Group FHIM is same as an expanded non-overlapped Group GHIM, i.e. the OvGroup FHIM optimization can be solved by an equivalent expanded Group FHIM optimization problem. This result is used in the implementation of OvGroup FHIM for our experiments.

6 Experiments

We use synthetic and real datasets to demonstrate the performance of our Group FHIM and OvGroup FHIM models, and compare it with LASSO [19], Hierarchical LASSO [1], Group Lasso [21], Trace-norm [9], Dirty model [8], QUIRE [15] and FHIM [16]. We use 80% of dataset for training and 20% for test, and 20% of training data as validation set to find optimal tuning parameters. We search tuning parameters for all methods using grid search, and for our model the parameters λ_β and λ_{a_k} are searched in the range of $[0.01, 100]$. In this paper, report our results on 5 simulations. Initialization, warm start, and stopping criterion play an important role for our Greedy Alternating Optimization algorithm mentioned in Algorithm 1. Below, we discuss how we choose them for our optimization. From our extensive experimental studies, we found that initializing $\boldsymbol{a}_k$ with $\mathbf{1}$ and $\boldsymbol{\beta}$ with $\boldsymbol{\beta}_{LASSO}$ works well for convergence.

6.1 Datasets We use synthetic datasets and 3 real datasets for classification and support recovery experiments.

6.1.1 Synthetic Dataset We generate the features of design matrix $\mathbf{X}$ using a normal distribution with mean zero and variance one ($\mathcal{N}(0,1)$). $\boldsymbol{\beta}, \boldsymbol{a_k}$ were generated as s-sparse vector from $\mathcal{N}(0,1)$, s is chosen as 5-10% of p and the number of groups $|\mathcal{G}| \in [10, 50]$. The group interaction weight matrix $\mathbf{W}_{OD}$ was generated using equation (4.7) for a $K \in [1, 5]$. The response vectors $\mathbf{y}$ was generated for logistic and large-margin formulation with a noise factor of 0.01. We generated several synthetic datasets by varying $n, p, K, |\mathcal{G}|$ and s. Note, we denote the combined total features (that is main effects + pairwise interaction) by q, here $q = p(p+1)/2$. In this paper, we show results for synthetic data in these settings: Case 1) $n > p$ and $q > n$ (high-dimensional setting w.r.t interaction features) and Case 2) $p > n$ (high-dimensional setting w.r.t original features).

6.1.2 Real Datasets To assess the performance of our model, we tested our methods on three prediction tasks:

1. *Classification on RCC sample*: This dataset contains 213 RCC samples from Benign and 4 different stages of tumor. Expression levels of 1092 proteins are collected in this dataset and these 1092 proteins belong to the 341 groups (overlapping groups). The number of Benign, Stage 1, Stage 2, Stage 3 and Stage 4 tumor samples are $40, 101, 17, 24$ and 31 respectively.
2. *Gene Expression Prediction*: This dataset [2] has 157 ChIP-Seq signals for transcription factor bindings and chromatin modifications and 1000 samples for gene transcripts. The features were grouped into 101 non-overlapping groups based on prior knowledge about ChIP-Seq experimental setup. For example, different ChIP-Seq experiments under different conditions or treatments for the same transcription factor are grouped into the same group.
3. *Peptide-MHC I Binding Prediction*: This dataset [10] is listed in Table 1. There are 9 positional groups (non-overlapping) in this dataset. Each positional group contains 20 features which are substitution log-odds from BLOSUM62 for the amino acid at this position.

Remark. RCC dataset was requested from the authors of [15]. ChIP-Seq data is publicly available at http://genome.ucsc.edu/ENCODE/downloads.html. Peptide-MHC I Binding dataset consists of publicly available data from Immune Epitope Database and Analysis Resource (IEDB) [20] which was used for training and privately collected data by our research collaborators which was used for testing.

6.2 Experimental Design and Evaluation metrics For synthetic data, we evaluate performance of our methods using prediction error and support recovery experiments. For real dataset, we perform the following evaluations:

1. RCC Classification: We perform 3 stage-wise binary classification experiments using RCC samples:
   (a) Case 1: Benign samples vs. Stage $1-4$.
   (b) Case 2: Benign and Stage 1 vs. Stage $2-4$.
   (c) Case 3: Benign, Stage $1, 2$ vs. Stage $3, 4$.
2. ChIP-Seq Gene Expression Classification: We perform two binary classification experiments: Case 1) predict gene expression levels as low or high, Case 2) predict whether genes are expressed or not.
3. Peptide-MHC I Binding Prediction: We predict binding peptides from non-binding peptides for three alleles, HLA-A*0201, HLA-A*0206 and HLA-A*2402.

For evaluation metrics, we use 1) F1-measure for support recovery of W_{OD} (synthetic) and 2) Area under ROC curve (ROC) for the classification (synthetic and real data).

Dataset	**#Peptides**	**#Binders**	**#Non-binders**
A0201-IEDB	8471	3939	8532
A0201-Japanese	114	59	55
A0206-IEDB	1820	951	869
A0206-Japanese	81	33	48
A2402-IEDB	2011	890	1121
A2402-Japanese	167	125	42

Table 1: Peptide-MHC I binding datasets

6.3 Performance on Synthetic dataset Tables 2 and 3 show that our Group FHIM and OvGroup FHIM outperforms the state-of-the-art approaches such as ℓ_1 Logistic Regression, Group Lasso [21], Hierarchical Lasso [1] and FHIM [16]. These models (except ℓ_1 Logistic Regression) were chosen for comparison because they are the state-of-the-art approaches which can recover grouping structure or high order feature interac-

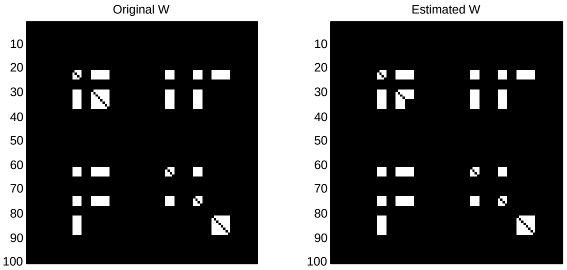

Figure 1: Support Recovery of $\mathbf{W}_{OD}$ (95 % sparse) for synthetic data $q = 5100, n = 1000$.

tions. Figure 1 shows an example for the support recovery of $\mathbf{W}_{OD}$ for the $q > n$ setting. From this figure, we see that our model performs very well (i.e. F_1 score is close to 1). For $p > n$ settings, our model also performs fairly well in the support recovery of $\mathbf{W}_{OD}$.

	ℓ_1 Logistic Reg.	**Group Lasso**	**Hier. Lasso**	**FHIM**	**Group FHIM (Log. Loss)**
$q > n$	0.52	0.74	0.58	0.89	**0.97**
$p > n$	0.51	0.52	-	0.54	**0.62**

Table 2: ROC scores on synthetic data with non-overlapping groups: case 1) q = 5100, n = 1000; case 2) p = 250, n = 100. Note: Hier. Lasso has heavy computation burden for $p > n$.

	ℓ_1 Logistic Reg.	**Overlap Group Lasso**	**Hier. Lasso**	**FHIM**	**OvGroup FHIM (Hinge Loss)**
$q > n$	0.54	0.67	0.56	0.69	**0.81**
$p > n$	0.53	0.58	-	0.57	**0.64**

Table 3: ROC scores on synthetic data with overlapping groups: case 1) q = 5,100, n = 1,000; case 2) p = 250, n = 100.

6.4 Classification Performance on RCC samples In this section, we report systematic experimental results on classification of samples from different stages of RCC. This dataset does not have grouping information for proteins. In order to group the proteins, we use the web based tool Database for Annotation, Visualization, and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/). There are a set of parameters that can be adjusted in DAVID based on which the functional classification is done. This whole set of parameters is controlled by a higher level parameter -"Classification Stringency", which determines how tight the resulting groups are in terms of association of the genes in each group. We set the stringency level to Medium which results in balanced functional groups where the association of the genes are moderately tight. The total number of groups based on cellular component annotations for RCC is 56. Each ungrouped gene forms a separate group, and in total we have 341 overlapping groups.

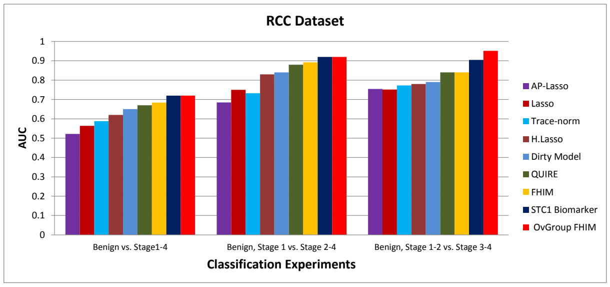

Figure 2: Comparison of the classification performance of different feature selection approaches with our model (OvGroup FHIM) in identifying the different stages of RCC.

The predictive performance of the bio-markers and pairwise group interactions selected by our OvGroup FHIM model (Hinge Loss) is compared against the markers selected by Lasso, All-Pairs Lasso [1], Group Lasso, Dirty model [8], QUIRE and FHIM. We use SLEP [13], MALSAR [22] packages for the implementation of most of these models. QUIRE and FHIM codes were obtained from the authors. The overall performance of the algorithms are shown in Figure 2. In this figure, we report average AUC score for five runs of 5-fold cross validation experiments for cancer stage prediction in RCC. The average ROC scores achieved by feature groups selected with our model are 0.72, 0.93 and 0.95 respectively for the three cases discussed in section 6.2. We performed pairwise t-tests for the comparisons of our method vs. the other methods, and all p-values were below 0.0075 which shows that our results are statistically significant. From Figure 2, we see that our model outperforms all the other algorithms for the three classification cases of RCC prediction and performs similarly to the well-known biomarker STC1. Interestingly, our OvGroup FHIM did not find any feature group interactions, i.e $\boldsymbol{a}_k = 0$ for the RCC dataset, and the feature groups (of $\boldsymbol{\beta}_g$) found by our model corresponds to the two groups containing STC1.

6.5 Gene Expression Prediction from ChIP-Seq Signals For case 1, the gene expression measured by Cap Analysis (CAGE) from the ENCODE project [2] above 3.0 (the median of nonzero gene expression levels) is considered as high, while the gene expression between 0 and 3.0 is considered as low for the classification experiments; for case 2, the genes with nonzero expression levels are considered as expressed and the others as non-expressed. Table 4 shows the gene expression prediction results on these two classification experiments. We observed that our Group FHIM outperforms all the state-

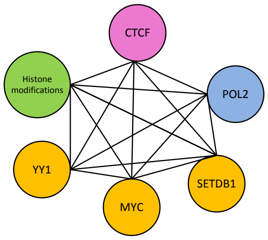

Figure 3: Interpretable interactions identified by OvGroup FHIM for predicting gene expression from ChIP-Seq signals.

	ℓ_1 Logistic Reg.	Group ℓ_1 Log. Reg.	FHIM	Group FHIM
Case 1	0.74	0.90	0.82	**0.92**
Case 2	0.72	0.89	0.80	**0.91**

Table 4: Gene Expression Prediction from ChIP-Seq signals

of-the-art models such as Group ℓ_1 logistic regression and FHIM. Moreover, our model discovers biologically meaningful ChIP-Seq signal interactions which are discussed in the section 6.5.1.

6.5.1 Feature Group Interactions discovered by Group FHIM An investigation of the interactions identified by our Group FHIM on the ChIP-Seq dataset reveals that many of these interactions are indeed relevant for gene expression. Figure 3 shows 6 out of the top 7 group interactions for the Case 1 classification, i.e., predicting whether a gene transcript is highly expressed or not. Among these group interactions, POL2 catalyzes DNA transcription and synthesizes mRNAs and most of small non-coding RNAs, and many transcription factors require its binding to gene promoters to begin gene transcription; MYC is known to recruit histone modifications to activate gene expression; YY1 is known to interact with histone modifications to activate or repress gene expression; SETDB1 regulates histone modifications to repress gene expression; CTCF is an insulator, its binding to MYC locus prevents the expression of MYC to be altered by DNA methylation, and it regulates chromatin structure for which its group also appeared in the dicriminative ones identified by our model. Further investigations of the interactions identified by our Group FHIM model might reveal novel insights that will help us to better understand gene regulation.

6.6 Peptide-MHC I Binding Prediction Table 5 shows the comparison of peptide-MHC I binding prediction of our model with respect to the state-of-the-art ℓ_1 and Group ℓ_1 logistic regression and FHIM. Figure 5 shows the ROC curves of Group FHIM and Group ℓ_1 logistic regression for Allele 0206. As evident from the AUC scores and ROC curve plots, our method achieves significant improvement over Group ℓ_1 logistic regression in separating the 'binders' from 'non-binders'. We found that ℓ_1 logistic regression gave slightly better performance on A2402, but our model identified meaningful group interactions as discussed below. Group ℓ_1 logistic regression produces worse performance than ℓ_1 logistic regression, which shows that only using grouping information does not help to identify discriminative individual features. However, our model Group FHIM significantly outperforms FHIM, which demonstrates the effectiveness of modeling both grouping information and high-order feature interactions.

Alleles	ℓ_1 Logistic Reg.	Group ℓ_1 Log. Reg.	FHIM	Group FHIM
A0201	0.74	0.72	0.72	**0.80**
A0206	0.76	0.75	0.68	**0.79**
A2402	**0.83**	0.77	0.75	0.82

Table 5: Peptide-MHC I binding prediction AUC scores

Figure 4 shows the factorized rank-1 interaction weight vector with absolute values greater than 0.1. This feature shows that the positions 2,5,6,9 interact; and moreover the interaction between the middle position and the position 9 is very important for predicting 9-mer peptide binding, which has experimental support from the crystal structure of the interaction complex [4]. We also found positions 2 and 9 interact for Alleles A0201 and A0206.

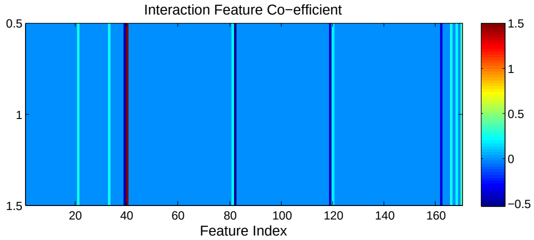

Figure 4: Interaction feature factor coefficients for A2402

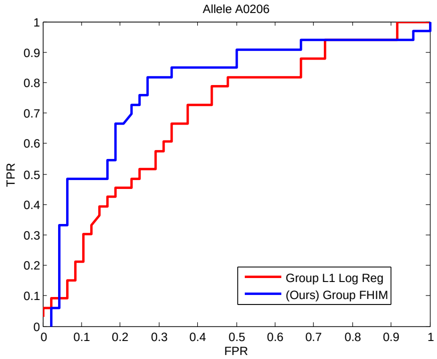

Figure 5: ROC curves for A0206

6.7 Computational Time Analysis

Group FHIM takes more time for convergence than the state-of-the-art approaches (LASSO and ℓ_1 logistic regression) since we do multiple rounds of greedy alternating optimization for $\boldsymbol{\beta}$ and $\boldsymbol{a}_k$. For $q > n$ setting with $n = 1000, p = 100, q = 5100, |\mathcal{G}| = 25$, our optimization method on Matlab takes around ~ 5 minutes to converge for fixed parameter, while for $p > n$ with $p = 250$, $n = 100$, our Group FHIM model takes around ~ 10 mins to converge. Our experiments were run on intel i3 dual-core 2.9GHz CPU with 8 GB RAM.

7 Conclusions

In this paper, we proposed a knowledge-based sparse learning framework called Group FHIM for identifying discriminative high-order feature group interactions in logistic regression and large-margin models, and studied interesting theoretical properties of our model. Empirical experiments on synthetic and real datasets showed that our model outperforms several well-known and state-of-the-art sparse learning techniques such as Lasso, ℓ_1 Logistic Regression, Group Lasso, Hierarchical Lasso, and FHIM, and it achieves comparable or better performance compared to the state-of-the-art knowledge based approaches such as QUIRE. Our model identifies high-order positional group interactions for peptide-MHC I binding prediction, and it discovers the important group interactions such as POL2-MYC, YY1-histone modifications, MYC-histone modifications, and CTCF-MYC which are valuable for understanding gene transcriptional regulation.

For future work, we will consider the following directions: (i) We will consider factorization of the weight matrix $\mathbf{W}$ as $\mathbf{W} = \sum_k \boldsymbol{a}_k \boldsymbol{b}_k^T$ since it is more general and can capture non-symmetric $\mathbf{W}$, (ii) For theoretical analysis, we will prove the Sparsistency assumption ($P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0) \to 1$) of theorem 5.2, and also study the asymptotic oracle properties for $p_n \to \infty$ as $n \to \infty$.

References

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

[2] C. Cheng, R. Alexander, R. Min, J. Leng, K. Y. Yip, J. Rozowsky, K.-K. Yan, X. Dong, S. Djebali, Y. Ruan, et al. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research*, 2012.

[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] D. K. Cole, P. J. Rizkallah, F. Gao, N. I. Watson, J. M. Boulter, J. I. Bell, M. Sami, G. F. Gao, and B. K. Jakobsen. Crystal structure of hla-a* 2402 complexed with a telomerase peptide. *European journal of immunology*, 2006.

[5] D. K. Duvenaud, H. Nickisch, and C. E. Rasmussen. Additive gaussian processes. In *NIPS*, pages 226–234, 2011.

[6] R. Foygel, N. Srebro, and R. Salakhutdinov. Matrix reconstruction with the local max norm. *arXiv preprint arXiv:1210.5196*, 2012.

[7] L. Jacob, G. Obozinski, and J.-P. Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, 2009.

[8] A. Jalali, P. Ravikumar, and S. Sanghavi. A dirty model for multiple sparse regression. *arXiv preprint arXiv:1106.5826*, 2011.

[9] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 457–464. ACM, 2009.

[10] P. P. Kuksa, M. R. Min, R. Dugar, and M. Gerstein. High-order neural networks and kernel methods for peptide-mhc binding prediction. *NIPS 2014 Workshop on Machine Learning in Computational Biology*, 2014.

[11] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *J. Mach. Learn. Res.*, 5:27–72, Dec. 2004.

[12] E. L. Lehmann and G. Casella. *Theory of point estimation*, volume 31. Springer, 1998.

[13] J. Liu, S. Ji, and J. Ye. *SLEP: Sparse Learning with Efficient Projections*. Arizona State University, 2009.

[14] M. R. Min, X. Ning, C. Cheng, and M. Gerstein. Interpretable sparse high-order boltzmann machines. In *The 17th International Conference on Artificial Intelligence and Statistics*, 2014.

[15] R. Min, S. Chowdhury, Y. Qi, A. Stewart, and R. Ostroff. An integrated approach to blood-based cancer diagnosis and biomarker discovery. In *Proceedings of the Pacific Symposium on Biocomputing*, 2014.

[16] S. Purushotham, M. R. Min, C.-C. J. Kuo, and R. Ostroff. Factorized sparse learning models with interpretable high order feature interactions. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14. ACM, 2014.

[17] E. Richard, G. R. Obozinski, and J.-P. Vert. Tight convex relaxations for sparse matrix factorization. In *Advances in Neural Information Processing Systems*, pages 3284–3292, 2014.

[18] J. A. Suykens and J. Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

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

[20] R. Vita, L. Zarebski, J. A. Greenbaum, H. Emami, I. Hoof, N. Salimi, R. Damle, A. Sette, and B. Peters. The immune epitope database 2.0. *Nucleic acids research*, 2010.

[21] 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.

[22] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.

[23] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.