Factorized Sparse Learning Models with Interpretable High Order Feature Interactions

Sanjay Purushotham*
University of Southern California
Los Angeles, CA, USA
spurusho@usc.edu

Martin Renqiang Min†
NEC Labs America
Princeton, NJ, USA
renqiang@nec-labs.com

C.-C. Jay Kuo
University of Southern California
Los Angeles, CA, USA
cckuo@sipi.usc.edu

Rachel Ostroff‡
SomaLogic, Inc.
Boulder, CO, USA
rostroff@somalogic.com

ABSTRACT

Identifying interpretable discriminative high-order feature interactions given limited training data in high dimensions is challenging in both machine learning and data mining. In this paper, we propose a factorization based sparse learning framework termed FHIM for identifying high-order feature interactions in linear and logistic regression models, and study several optimization methods for solving them. Unlike previous sparse learning methods, our model FHIM recovers both the main effects and the interaction terms accurately without imposing tree-structured hierarchical constraints. Furthermore, we show that FHIM has oracle properties when extended to generalized linear regression models with pairwise interactions. Experiments on simulated data show that FHIM outperforms the state-of-the-art sparse learning techniques. Further experiments on our experimentally generated data from patient blood samples using a novel SOMAmer (Slow Off-rate Modified Aptamer) technology show that, FHIM performs blood-based cancer diagnosis and bio-marker discovery for Renal Cell Carcinoma much better than other competing methods, and it identifies interpretable block-wise high-order gene interactions predictive of cancer stages of samples. A literature survey shows that the interactions identified by FHIM play important roles in cancer development.

Categories and Subject Descriptors

J.3 [**Computer Applications**]: Life and Medical Sciences; I.2.6 [**Computing Methodologies**]: Artificial Intelligence—*sparse learning, feature selection*

Keywords

Sparse Learning; High-order interactions; Biomaker Discovery; Blood-based Cancer Diagnosis

1. INTRODUCTION

Identifying interpretable high-order feature interactions is an important problem in machine learning, data mining, and biomedical informatics, because feature interactions often help reveal some hidden domain knowledge and the structures of problems under consideration. For example, genes and proteins seldom perform their functions independently, so many human diseases are often manifested as the dysfunction of some pathways or functional gene modules, and the disrupted patterns due to diseases are often more obvious at a pathway or module level. Identifying these disrupted gene interactions for different diseases such as cancer will help us understand the underlying mechanisms of the diseases and develop effective drugs to cure them. However, identifying reliable discriminative high-order gene/protein or SNP interactions for accurate disease diagnosis such as early cancer diagnosis directly based on patient blood samples is still a challenging problem, because we often have very limited patient samples but a huge number of complex feature interactions to consider.

In this paper, we propose a sparse learning framework based on weight matrix factorizations and ℓ_1 regularizations for identifying discriminative high-order feature interactions in linear and logistic regression models, and we study several optimization methods for solving them. 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 interactions for disease status prediction. Our proposed sparse learning framework is 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 single

*Co-first author
†Co-first author, corresponding author
‡To whom data request should be sent

KDD'14, August 24–27, 2014, New York, NY, USA.

http://dx.doi.org/10.1145/2623330.2623747.

discriminative features and discriminative block-wise high-order interactions in a sparse learning framework, which can be easily extended to handle arbitrarily high-order feature interactions; (2) Our method works on high-dimensional input feature spaces and ill-posed problems with much more features than data points, which is typical for biomedical applications such as biomarker discovery and cancer diagnosis; (3) Our method has interesting theoretical properties for generalized linear regression models; (4) The interactions identified by our method lead to biomedical insight into understanding blood-based cancer diagnosis.

2. RELATED WORK

Variable selection has been a well studied topic in statistics, machine learning, and data mining literature. Generally, variable selection approaches focus on identifying discriminative features using regularization techniques. Most recent methods focus on identifying discriminative features or groups of discriminative features based on Lasso penalty [18], Group Lasso [21], Trace-norm [6], Dirty model [8] and Support Vector Machines (SVMs) [16]. A recent approach [20] heuristically adds some possible high-order interactions into the input feature set in a greedy way based on lasso penalized logistic regression. Some recent approaches [2],[3] enforce strong and/or weak heredity constraints to recover the pairwise interactions in linear regression models. In strong heredity, an interaction term can be included in the model only if the corresponding main terms are also included in the model, while in weak heredity, an interaction term is included when either of the main terms are included in the model. However, recent studies in bioinformatics has shown that feature interactions need not follow heredity constraints for manifestation of the diseases, and thus the above approaches [2],[3] have limited chance of recovering relevant interactions. Kernel methods such as Gaussian Process [4] and Multiple Kernel Learning [10] can be used to model high-order feature interactions, but they can only tell which orders are important. Thus, all these previous approaches either failed to identify specific high-order interactions for prediction or identified sporadic pairwise interactions in a greedy way, which is very unlikely to recover the 'interpretable' blockwise high-order interactions among features in different sub-components (for example: pathways or gene functional modules) of systems. Recently, [14] proposed an efficient way to identify combinatorial interactions among interactive genes in complex diseases by using overlapping group lasso and screening. However, they use prior information such as gene ontology in their approach, which is generally not available or difficult to collect for some machine learning problems. Thus, there is a need to develop new efficient techniques to automatically capture the important 'blockwise' high-order feature interactions in regression models, which is the focus of this paper.

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 our approach, and in section 5 we give a overview of theoretical properties associated with our method. In section 6, we present the optimization methods which we use to solve our optimization problem. In section 7, we discuss our experimental setup and present our results on synthetic and real datasets. Finally, in section 8 we conclude the paper with discussions and future research directions.

3. PROBLEM FORMULATION

Consider a regression setup with a training set of n samples and p features, $\{(\mathbf{X}^{(i)}, y^{(i)})\}$, where $\mathbf{X}^{(i)} \in \mathbb{R}^p$ is the i^{th} instance (column) of the design matrix $\mathbf{X}$ $(p \times n)$, $y^{(i)} \in \mathbb{R}$ is the i^{th} instance of response variable $\mathbf{y}$ $(n \times 1)$, and $i = 1, \ldots, n$. To model the response in terms of the predictors, we can set up a linear regression model

$$y^{(i)} = \boldsymbol{\beta}^T \mathbf{X}^{(i)} + \epsilon^{(i)}, \tag{1}$$

or a logistic regression model

$$p(y^{(i)} = 1 | \mathbf{X}^{(i)}) = \frac{1}{1 + \exp(-\boldsymbol{\beta}^T \mathbf{X}^{(i)} - \beta_0)}, \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the weight vector associated with single features (also called main effects), $\boldsymbol{\epsilon} \in \mathbb{R}^n$ is a noise vector, and $\beta_0 \in \mathbb{R}$ is the bias term. In many practical fields such as bioinformatics and medical informatics, the main terms (the terms only involving single features) are not enough to capture complex relationship between the response and the predictors, and thus high-order interactions are necessary. In this paper, we consider regression models with both main effects and high-order interaction terms. Equation 3 shows a linear regression model with pairwise interaction terms.

$$y^{(i)} = \boldsymbol{\beta}^T \mathbf{X}^{(i)} + \mathbf{X}^{(i)T} \mathbf{W} \mathbf{X}^{(i)} + \epsilon^{(i)}, \tag{3}$$

where $\mathbf{W}(p \times p)$ is the weight matrix associated with the pairwise feature interactions. The corresponding loss function (the sum of squared errors) is as follows (we center the data to avoid an additional bias term),

$$L_{sqerr}(\boldsymbol{\beta}, \mathbf{W}) = \frac{1}{2} \sum_{i=1}^{n} ||y^{(i)} - \boldsymbol{\beta}^T \mathbf{X}^{(i)} - \mathbf{X}^{T(i)} \mathbf{W} \mathbf{X}^{(i)}||_2^2. \tag{4}$$

We can similarly write the logistic regression model with pairwise interactions as follows,

$$p(y^{(i)} | \mathbf{X}^{(i)}) = \frac{1}{1 + \exp(-y^{(i)}(\boldsymbol{\beta}^T \mathbf{X}^{(i)} + \mathbf{X}^{(i)T} \mathbf{W} \mathbf{X}^{(i)} + \beta_0))} \tag{5}$$

and the corresponding loss function (the sum of the negative log-likelihood of the training data) is,

$$\begin{aligned} L_{logistic}(\boldsymbol{\beta}, \mathbf{W}, \beta_0) &= \sum_{i=1}^{n} \log(1 + \exp(-y^{(i)}(\boldsymbol{\beta}^T \mathbf{X}^{(i)} \\ &\quad + \mathbf{X}^{(i)T} \mathbf{W} \mathbf{X}^{(i)} + \beta_0)). \end{aligned} \tag{6}$$

4. OUR APPROACH

In this section, we propose an optimization-driven sparse learning framework to identify discriminative single features and groups of high-order interactions among input features for output prediction in the setting of limited training data. When the number of input features is huge (e.g. biomedical applications), it is practically impossible to explicitly consider quadratic or even higher-order interactions among all the input features based on simple lasso penalized linear regression or logistic regression. To solve this problem, we propose to factorize the weight matrix $\mathbf{W}$ associated with high-order interactions between input features to be a sum of K rank-one matrices for pairwise interactions or a sum of low-rank high-order tensors for higher-order interactions.

Each rank-one matrix for pairwise feature interactions is represented by an outer product of two identical vectors, and each m-order ($m > 2$) tensor is represented by the outer product of m identical vectors. Besides minimizing the loss function of linear regression or logistic regression, we penalize the ℓ_1 norm of both the weights associated with single input features and the weights associated with high-order feature interactions. Mathematically, we solve the optimization problem to identify the discriminative single and pairwise interaction features as follows,

$$\{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{a}}_k\} = \arg\min_{\boldsymbol{a}_k, \boldsymbol{\beta}} L_{sqerr}(\boldsymbol{\beta}, \mathbf{W}) + \lambda_\beta \|\boldsymbol{\beta}\|_1 + \sum_{k=1}^{K} \lambda_{a_k} \|\boldsymbol{a}_k\|_1 \tag{7}$$

where $\mathbf{W} = \sum_{k=1}^{K} \boldsymbol{a}_k \odot \boldsymbol{a}_k$, $\odot$ represents the tensor product/outer product, and $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{a}}_k$ represent the estimated parameters of our model and let Q represent objective function of (7) . For logistic regression, we replace $L_{sqerr}(\boldsymbol{\beta}, \mathbf{W})$ in (7) by $L_{logistic}(\boldsymbol{\beta}, \mathbf{W}, \beta_0)$. We call our model Factorization-based High-order Interaction Model (FHIM).

PROPOSITION 4.1. *The optimization problem in Equation 7 is convex in $\boldsymbol{\beta}$ and non-convex in $\boldsymbol{a}_k$.*

Because of the non-convexity property of our optimization problem, it is difficult to propose optimization algorithms which guarantee convergence to global optima. Here, we adopt a greedy alternating optimization methods to find the local optima for our problem. In the case of pairwise interactions, fixing other weights, we solve each rank-one weight matrix each time. Please note that our symmetric positive definite factorization of $\mathbf{W}$ makes this sub-optimization problem very easy. Moreover, for a particular rank-one weight matrix $\boldsymbol{a}_k \odot \boldsymbol{a}_k$, the nonzero entries of the corresponding vector $\boldsymbol{a}_k$ can be interpreted as the block-wise interaction feature indices of a densely interacting feature group. In the case of higher-order interactions, the optimization procedure is similar to the one for the pairwise interactions except that we have more rounds of alternating optimization. The parameter K of $\mathbf{W}$ is generally unknown in real datasets, thus, we greedily estimate K during the alternating optimization algorithm. In fact, the combination of our factorization formulation and the greedy algorithm is effective for estimating the interaction weight matrix $\mathbf{W}$. $\boldsymbol{\beta}$ is re-estimated when K is greedily added during the alternating optimization as shown in algorithm 1.

Algorithm 1 Greedy Alternating Optimization

```
Initialize β to 0, K = 1 and a_K = 1
While (K==1) || (a_{K-1} ≠ 0 for K > 1)
    Repeat until convergence
        β_j^t = arg min_j Q(β_1^t, ..., β_{j-1}^t, β_{j+1}^{t-1}, β_p^{t-1}), a_k^{t-1})
        a_{k,j}^t = arg min_j Q((a_{k,1}^t, ..., a_{k,j-1}^t, a_{k,j+1}^{t-1}, a_{k,p}^{t-1}), β^t)
    End Repeat
    K = K + 1; a_K = 1
End While
Remove a_K and a_{K-1} from a.
```

5. THEORETICAL PROPERTIES

In this section, we study the asymptotic behavior of FHIM for the likelihood-based generalized linear regression models. The lemmas and theorems proved here are similar to the ones shown in the paper [3]. However, in their paper the authors make an assumption on the strong heredity (i.e. interaction term coefficients are dependent on the main effects), which is not assumed in our model since we are interested in identifying all high-order interactions irrespective of heredity constraints. Here, we discuss the asymptotic properties w.r.t to the main effects and factorized co-efficients.

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 $f(g(\mathbf{X}_i), y_i)$ conditioned on $\mathbf{X}_i$, where g is a known regression function with main effects and all possible pairwise interactions. Let β_j^* and $a_{k,j}^*$ denote the underlying true parameters satisfying block-wise properties implied by our factorization. Let $Q_n(\boldsymbol{\theta})$ denote the objective with negative log-likelihood and $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*T}, \boldsymbol{\alpha}^{*T})^T$, where $\boldsymbol{\alpha}^* = (\boldsymbol{a}_k^*), k = 1, ..., K$. We consider the estimates for FHIM as $\hat{\boldsymbol{\theta}}_n$:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_n &= \arg\min_{\theta} Q_n(\boldsymbol{\theta}) \\ &= \arg\min_{\theta} -\frac{1}{n}\sum_{i=1}^{n}(L(g(\mathbf{X}_i), y_i) + \lambda_\beta|\boldsymbol{\beta}| + \sum_k \lambda_{\alpha_k}|\boldsymbol{\alpha_k}|\end{aligned} \tag{8}$$

where $L(g(\mathbf{X}_i), y_i)$ is the loss function of generalized linear regression models with pairwise interactions. In the case of linear regression, $g(\cdot)$ takes the form of Equation (3) without the noise term ϵ and $L(\cdot)$ takes the form of Equation (4). Now, let us define

$$\begin{aligned}&\mathcal{A}_1 = \{j : \beta_j^* \neq 0\} \\ &\mathcal{A}_2 = \{(k,l) : \alpha_{k,l}^* \neq 0\}, \\ &\mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2\end{aligned} \tag{9}$$

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

$$\begin{aligned}a_n &= \max\{\lambda_j^\beta, \lambda_l^{\alpha_k} : j \in \mathcal{A}_1, (k,l) \in \mathcal{A}_2\} \\ b_n &= \min\{\lambda_j^\beta, \lambda_l^{\alpha_k} : j \in \mathcal{A}_1^c, (k,l) \in \mathcal{A}_2^c\}\end{aligned} \tag{10}$$

Now, we show that our model possesses the oracle properties for (i) $n \to \infty$ with fixed p and (ii) $p_n \to \infty$ as $n \to \infty$ under some regularity conditions. Please refer to Appendix for proofs of the lemmas & theorems of sections 5.1 and 5.2.

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 lemmas and theorems. FHIM possesses oracle properties [3] under certain regularity conditions (C1)-(C3) shown below. Let Ω denote the parameter space for $\boldsymbol{\theta}$.

(C1) 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

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

(C2) 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 $\boldsymbol{\theta} = \boldsymbol{\theta}^*$.

(C3) 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

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

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 [11].

LEMMA 5.1. *Assume $a_n = o(1)$ as $n\to\infty$. Then under regularity conditions (C1)-C3), there exists a local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*|| = O_P(n^{-1/2}+a_n)$*

THEOREM 5.2. *Assume $\sqrt{n}b_n \to \infty$ and the minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ given in lemma 5.1 satisfies $||\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*|| = O_P(n^{-1/2})$. Then under regularity conditions (C1)-(C3), we have*

$$P(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1^C} = 0) \to 1, \qquad P(\hat{\boldsymbol{\alpha}}_{\mathcal{A}_2^C} = 0) \to 1$$

Lemma 5.1 implies that when the tuning parameters associated with the non-zero coefficients of main effects and 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 shows that our model removes noise consistently with high probability ($\to 1$). If $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$, then lemma 5.1 and theorem 5.2 imply that the $\sqrt{n}-$consistent estimator $\hat{\boldsymbol{\theta}}_n$ satisfies $P(\hat{\boldsymbol{\theta}}_{\mathcal{A}^c} = 0) \to 1$.

THEOREM 5.3. *Assume $\sqrt{n}a_n \to 0$ and $\sqrt{n}b_n \to \infty$. Then under the regularity conditions (C1)-(C3), the component $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mathcal{A}}}$ of the local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ (given in lemma 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.

Theorem 5.3 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 5.2 and 5.3, we can say that our model has the oracle property [3], [5], 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 $\lambda_\beta, \lambda_{\alpha_k}$ (see appendix for more details). Thus, our tuning parameters are:

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

5.2 Asymptotic Oracle Properties When $p_n \to \infty$ as $n \to \infty$

In this section, we consider the asymptotic behavior of our model when the number of predictors p_n grows to infinity along with the sample size n. If certain regularity conditions (C4)-(C6) (shown below) hold, then we can show that our model possesses the oracle property.

We denote the total number of predictors by p_n. We denote all the quantities that change with sample size by adding n as their subscript. $\mathcal{A}_1, \mathcal{A}_2, \mathcal{A}$ are defined as in section 5 and let $s_n = |\mathcal{A}_n|$. The asymptotic properties of our model when the number of predictors increases along with the sample size are described in the following lemma and theorem. The regularity conditions (C4)-(C6) are given below: Let Ω_n denote the parameter space for $\boldsymbol{\theta}_n$.

(C4) The observations $\mathbf{V}_{ni} : i = 1,...,n$ are independent and identically distributed with a probability density $f_n(\mathbf{V}_n, \boldsymbol{\theta}_n)$, which has a common support. We assume the density f_n satisfies the following equations:

$$E_{\boldsymbol{\theta}_n}\Big[\frac{\partial \log f_n(\mathbf{V}_n,\boldsymbol{\theta}_n)}{\partial\theta_{nj}}\Big] = 0 \qquad \text{for } j = 1,...,p_n,$$

and

$$\begin{aligned}\mathbf{I}_{jk}(\boldsymbol{\theta}_n) =& E_{\boldsymbol{\theta}_n}\Big[\frac{\partial \log f_n(\mathbf{V}_n,\boldsymbol{\theta}_n)}{\partial\theta_{nj}}\frac{\partial \log f_n(\mathbf{V}_n,\boldsymbol{\theta}_n)}{\partial\theta_{nk}}\Big]\\ =& E_{\boldsymbol{\theta}_n}\Big[-\frac{\partial^2 \log f_n(\mathbf{V}_n,\boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nk}}\Big]\end{aligned}$$

(C5) $I_n(\boldsymbol{\theta}_n) = E[(\frac{\partial \log f_n(\mathbf{V}_{n1},\boldsymbol{\theta}_n)}{\partial\boldsymbol{\theta}_n})(\frac{\partial \log f_n(\mathbf{V}_{n1},\boldsymbol{\theta}_n)}{\partial\boldsymbol{\theta}_n})^T]$ satisfies
$0 < C_1 < \lambda_{\min}I_n(\boldsymbol{\theta}_n) \le \lambda_{\max}I_n(\boldsymbol{\theta}_n) < C_2 < \infty$ for all n, where $\lambda_{\min}(.)$ and $\lambda_{\max}(.)$ represent the smallest and largest eigenvalues of a matrix respectively. Moreover, for any $j,k = 1,...,p_n$,

$$\begin{aligned}&E_{\boldsymbol{\theta}_n}\left\{\frac{\partial \log f_n(\mathbf{V}_{n1},\boldsymbol{\theta}_n)}{\partial\theta_{nj}}\frac{\partial \log f_n(\mathbf{V}_{n1},\boldsymbol{\theta}_n)}{\partial\theta_{nk}}\right\}^2\\ &< C_3 < \infty,\end{aligned}$$

and

$$E_{\boldsymbol{\theta}_n}\left\{\frac{\partial^2 \log f_n(\mathbf{V}_{n1},\boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nk}}\right\} < C_4 < \infty$$

(C6) There exists a large open set $\omega_n \subset \Omega_n \in \mathbb{R}^{p_n}$ which contains the true parameters $\boldsymbol{\theta}^*_n$ such that for almost all $\mathbf{V}_{ni}$ the density admits all third derivatives $\partial^3 f_n(\mathbf{V}_{ni},\boldsymbol{\theta}_n)/\partial\theta_{nj}\,\partial\theta_{nk}\partial\theta_{nl}$ for all $\boldsymbol{\theta}_n \in \omega_n$. Furthermore, there are functions M_{njkl} such that

$$\left|\frac{\partial^3 f_n(\mathbf{V}_{ni},\boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nk}\partial\theta_{nl}}\right| \le M_{njkl}(\mathbf{V}_{ni})$$

for all $\boldsymbol{\theta}_n \in \omega_n$ and

$$E_{\boldsymbol{\theta}_n}M^2_{njkl}(\mathbf{V}_{ni}) < C_5 < \infty$$

for all p_n, n, and j,k,l.

LEMMA 5.4. *Assume that the density $f_n(\boldsymbol{V_n}, \boldsymbol{\theta^*_n})$ satisfies some regularity conditions (C4)-(C6). If $\sqrt{n}a_n \to 0$ and $p_n^5/n \to 0$ as $n\to\infty$, then there exists a local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \boldsymbol{\theta}^*_{\boldsymbol{n}}|| = O_P(\sqrt{p_n}(n^{-1/2}+a_n))$*

THEOREM 5.5. *Suppose that the density $f_n(\boldsymbol{V_n}, \boldsymbol{\theta_n^*})$ satisfies some regularity conditions (C4)-(C6). If $\sqrt{np_n}a_n \to 0$, $\sqrt{n/p_n}b_n \to \infty$ and $p_n^5/n \to 0$ as $n \to \infty$, then with probability tending to 1, the $\sqrt{n/p_n}$-consistent local minimizer $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ in Lemma 5.4 satisfies the following:*

- *Sparsity:* $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n^c} = \mathbf{0}$
- *Asymptotic normality:*
$\sqrt{n}\boldsymbol{A}_n\boldsymbol{I}_n^{\frac{1}{2}}(\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} - \boldsymbol{\theta}^*_{n\mathcal{A}_n}) \to_d N(\mathbf{0}, \boldsymbol{G})$

where $\boldsymbol{A}_n$ is an arbitrary $m \times s_n$ matrix with finite m such that $\boldsymbol{A}_n\boldsymbol{A}_n^T \to \mathbf{G}$ and $\mathbf{G}$ is a $m \times m$ nonnegative symmetric matrix and $\mathbf{I}_n(\boldsymbol{\theta}^*_{n\mathcal{A}_n})$ is the Fisher information matrix of $\boldsymbol{\theta}_{n\mathcal{A}_n}$ at $\boldsymbol{\theta}_{n\mathcal{A}_n} = \boldsymbol{\theta}^*_{n\mathcal{A}_n}$. Since the dimension of $\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n} \to \infty$ as sample size $n \to \infty$, we could consider arbitrary linear combination $\mathbf{A}_n\hat{\boldsymbol{\theta}}_{n\mathcal{A}_n}$ for the asymptotic normality of our model's estimates. Similar to section 5.1, to satisfy oracle property, we have to consider an adaptive weights $w_{nj}^{\beta}, w_{nl}^{\alpha_k}$ [23] for our tuning parameters $\lambda_\beta, \lambda_{\alpha_k}$ as:

$$\lambda_{nj}^{\beta} = \frac{\log(n)p_n}{n}\lambda_\beta w_{nj}^{\beta}, \qquad \lambda_{n,l}^{\alpha_k} = \frac{\log(n)p_n}{n}\lambda_{\alpha_k} w_{nl}^{\alpha_k}$$

6. OPTIMIZATION

In this section, we outline three optimization methods that we employ to solve our objective function (7), which corresponds to Line 4 and 5 in Algorithm 1. [15] provides a good survey on several optimization approaches for solving ℓ_1-regularized regression problems. In this paper, we use the sub-gradient and co-ordinate wise soft-thresholding based optimization methods since they work well and are easy to implement. We compare these methods in the experimental results in section 7.

6.1 Sub-Gradient Methods

Sub-gradient based strategies treat the non-differentiable objective as a non-smooth optimization problem and use sub-gradients of the objective function at the non-differentiable points. For our model, the optimality conditions w.r.t parameter vectors $\boldsymbol{\beta}$ and $\boldsymbol{a}_k$ can be written out separately based on the objective function (7). Optimality conditions w.r.t $\boldsymbol{a}_k$ is:

$$\begin{cases} \nabla_j\mathcal{L}(\boldsymbol{a}_k) + \lambda_{a_k} sgn(a_{kj}) = 0 & |a_{kj}| > 0 \\ |\nabla_j\mathcal{L}(\boldsymbol{a}_k)| \le \lambda_{a_k} & a_{kj} = 0 \end{cases}$$

where $\mathcal{L}(\boldsymbol{a}_k)$ is the loss function of our linear regression model or logistic regression model in Equation (7) w.r.t $\boldsymbol{a}_k$. Similarly, optimality conditions can be written for $\boldsymbol{\beta}$. The sub-gradient $\nabla_j^s f(\boldsymbol{a}_k)$ for each a_{kj} is given by

$$\nabla_j^s f(\boldsymbol{a}_k) = \begin{cases} \nabla_j\mathcal{L}(\boldsymbol{a}_k) + \lambda_{a_k} sgn(a_{kj}), & |a_{kj}| > 0 \\ \nabla_j\mathcal{L}(\boldsymbol{a}_k) + \lambda_{a_k}, & a_{kj} = 0, \nabla_j\mathcal{L}(\boldsymbol{a}_k) < -\lambda_{a_k} \\ \nabla_j\mathcal{L}(\boldsymbol{a}_k) - \lambda_{a_k}, & a_{kj} = 0, \nabla_j\mathcal{L}(\boldsymbol{a}_k) > \lambda_{a_k} \\ 0, & -\lambda_{a_k} \le \nabla_j\mathcal{L}(\boldsymbol{a}_k) \le \lambda_{a_k} \end{cases}$$

where

$$\nabla_j\mathcal{L}(\boldsymbol{a}_k) = \frac{1}{2}\sum_i(-2\mathbf{X}_j^{(i)}\mathbf{X}^{T(i)}\boldsymbol{a}_k)[y^{(i)} - \boldsymbol{\beta}^T\mathbf{X}^{(i)} - \mathbf{X}^{T(i)}\mathbf{W}\mathbf{X}^{(i)}].$$

for our linear regression model. The negation of the sub-gradient represents the steepest descent direction. Similarly the sub-gradients for $\boldsymbol{\beta}$ ($\nabla_j^s f(\boldsymbol{\beta})$) can be calculated. Differential of the loss function of the linear regression in Equation (7) w.r.t $\boldsymbol{\beta}$ is given by

$$\nabla_j\mathcal{L}(\boldsymbol{\beta}) = \frac{1}{2}\sum_i(-2\mathbf{X}_j^{(i)})[y^{(i)} - \boldsymbol{\beta}^T\mathbf{X}^{(i)} - \mathbf{X}^{T(i)}\mathbf{W}\mathbf{X}^{(i)}]$$

6.1.1 Orthant-Wise Descent (OWD)

Andrew and Gao [1] proposed an effective strategy for solving large-scale ℓ_1-regularized regression problems based on choosing an appropriate steepest descent direction for the objective function and taking a step like a Newton iteration in this direction (with an L-BFGS Hessian approximation [12]). The orthant-wise learning descent method for our model takes the following form

$$\boldsymbol{\beta} \leftarrow \mathcal{P}_{\mathcal{O}}[\boldsymbol{\beta} - \gamma_{\boldsymbol{\beta}}\mathcal{P}_{\mathcal{S}}[H_{\boldsymbol{\beta}}^{-1}\nabla^s f(\boldsymbol{\beta})]]$$
$$\boldsymbol{a_k} \leftarrow \mathcal{P}_{\mathcal{O}}[\boldsymbol{a_k} - \gamma_{\boldsymbol{a_k}}\mathcal{P}_{\mathcal{S}}[H_{\boldsymbol{a_k}}^{-1}\nabla^s f(\boldsymbol{a_k})]]$$

where $\mathcal{P}_{\mathcal{O}}$ and $\mathcal{P}_{\mathcal{S}}$ are two projection operators and $H_{\boldsymbol{\beta}}$ is the positive definite approximation of Hessian of quadratic approximation of objective function $f(\boldsymbol{\beta})$, and $\gamma_{\boldsymbol{\beta}}$ and $\gamma_{\boldsymbol{a_k}}$ are step sizes. $\mathcal{P}_{\mathcal{S}}$ projects the Newton-like direction to guarantee that it is in the descent direction. $\mathcal{P}_{\mathcal{O}}$ projects the step onto the orthant containing $\boldsymbol{\beta}$ or $\boldsymbol{a}_k$ and ensures that line search does not cross points of non-differentiability.

6.1.2 Projected Scaled Sub-Gradient (PSS)

Schmidt [15] proposed optimization methods called Projected Scaled Sub-Gradient methods where the iterations can be written as the projection of a scaling of a sub-gradient of the objective. Please refer to [1] and [15] for more details on OWD and PSS methods.

6.2 Soft-thresholding

Soft-thresholding based co-ordinate descent optimization method can be used to find $\boldsymbol{\beta}, \boldsymbol{a}_k$ updates in the alternating optimization algorithm for our FHIM model. The $\boldsymbol{\beta}$ updates are $\tilde{\beta}_j$ and are given by

$$\tilde{\beta}_j(\lambda_\beta) \leftarrow S\Big(\tilde{\beta}_j(\lambda_\beta) + \sum_{i=1}^n X_{ij}(y_i - \sum_{k\neq j} X_{jk}\tilde{\beta}_k - \sum_k X_{ik}\mathbf{W}X_{ki}), \lambda_\beta\Big)$$

where $\mathbf{W} = \sum_k \boldsymbol{a_k} \odot \boldsymbol{a_k}$, and S is the soft-thresholding operator [7]. Similarly, the updates for $\boldsymbol{a_k}$ are $\tilde{a}_{kj}$ and given by

$$\tilde{a}_{kj}(\lambda_{a_k}) \leftarrow S\Big(\tilde{a}_{kj}(\lambda_{a_k}) + \sum_{i=1}^n X_{ij}(\sum_k\sum_{r=1}^p a_{kr}X_{ir})[y_i - \sum_{k\neq j} X_{jk}\tilde{\beta}_k - \sum_k X_{ik}\mathbf{W}_{\sim j}X_{ki}], \lambda_{a_k}\Big)$$

where $\mathbf{W}_{\sim j}$ is $\mathbf{W}$ with j^{th} column and j^{th} row elements are all zero.

7. EXPERIMENTS

In this section, we use synthetic and real datasets to demonstrate the efficacy of our model (FHIM), and compare its performance with LASSO [18], All-Pairs Lasso [2], Hierarchical LASSO [2], Group Lasso [21], Trace-norm [9], Dirty model [8] and QUIRE [14]. For all these models, we perform 5 runs of 5-fold cross-validation on training dataset (80 %)

to find the optimal parameters and evaluate prediction error on a test dataset (20 %). 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 $[0.01, 10]$. We also discuss the support recovery of $\boldsymbol{\beta}$ and $\mathbf{W}$ for our model.

7.1 Datasets

We use synthetic datasets and a real dataset for classification and support recovery experiments. We give detailed description of these datasets below.

7.1.1 Synthetic Dataset

We generate the predictors of the design matrix $\mathbf{X}$ using a normal distribution with mean zero and variance one. The weight matrix $\mathbf{W}$ was generated as a sum of K rank one matrices i.e. $\mathbf{W} = \sum_{k=1}^{K} \boldsymbol{a}_k \boldsymbol{a}_k^T$. $\boldsymbol{\beta}, \boldsymbol{a_k}$ were generated as a sparse vector from a normal distribution with mean 0 and variance 1, while noise vector $\boldsymbol{\epsilon}$ is generated from a normal distribution with mean 0 and variance 0.1. Finally, the response vectors $\mathbf{y}$ of the linear and logistic regression models with pairwise interactions were generated using Equations (3) and (5) respectively. We generated several synthetic datasets by varying number of instances (n), number of variables/predictors (p), rank of $\mathbf{W}$ i.e. K and sparsity level of $\boldsymbol{\beta}$, $\boldsymbol{a}_k$. We denote the combined total predictors (that is main effects predictors + predictors for interaction terms) by q, here $q = p(p+1)/2$. Sparsity level (non-zeros) was chosen as $2 \sim 4\%$ for large $p(> 100)$, and $5 \sim 10\%$ for small $p(< 100)$ for both $\boldsymbol{\beta}, \boldsymbol{a}_k$. 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 combined predictors) and, Case (2) $p > n$ (high-dimensional w.r.t original predictors).

7.1.2 Real Dataset

To predict cancer progression status directly from blood samples, we generated our own dataset. All samples and clinical information were collected under Health Insurance Portability and Accountability Act compliance from study participants after obtaining written informed consent under clinical research protocols approved by the institutional review boards for each site. Blood was processed within 2 hours of collection according to established standard operating procedures. To predict RCC status, serum samples were collected at a single study site from patients diagnosed with RCC or benign renal mass prior to treatment. Definitive pathology diagnosis of RCC and cancer stage was made after resection. Outcome data was obtained through follow-up from 3 months to 5 years after initial treatment. Our RCC dataset contains 212 RCC samples from benign and 4 different stages of tumor. Expression levels of 1092 proteins based on a high-throughput SOMAmer protein quantification technology are collected. The number of Benign, Stage 1, Stage 2, Stage 3 and Stage 4 tumor samples are 40; 101; 17; 24 and 31 respectively.

7.1.3 Experimental Design

We use linear regression models (Equation 3) for all the following experiments and we only use logistic regression models (Equation 5) for synthetic data experiments shown in table 2. We evaluate the performance of our method (FHIM) by the following experiments:

1. Prediction error and support recovery experiments on synthetic datasets
2. Classification experiments using RCC samples: We perform three stage-wise binary classification experiments using RCC samples:
   (a) Case 1: Classification of Benign samples from Stage $1-4$ samples.
   (b) Case 2: Classification of Benign and Stage 1 samples from Stage $2-4$ samples.
   (c) Case 3: Classification of Benign, Stage $1, 2$ samples from Stage $3, 4$ samples.

7.1.4 Performance on Synthetic dataset

We evaluate the performance of our model (FHIM) on synthetic dataset by the following experiments: (i) Comparison of optimization methods presented in section 6, (ii) Prediction error on the test data for $q > n$ and $p > n$ (high-dimensional settings), (iii) Support recovery accuracy of $\boldsymbol{\beta}$, $\mathbf{W}$ and (iv) Prediction of rank of $\mathbf{W}$ using greedy approach.

Table 3 shows the prediction error on test data when different optimization methods (discussed in section 6) are used for our model (FHIM). From table 3, we see that both OWD and PSS methods perform nearly similar (OWD is marginally better), and are better than the soft-thresholding method. This is because, in soft-thresholding, co-ordinate updates of variables might not be accurate in high dimensional settings (i.e. the solution is affected by the path taken during updates). We observed that soft-thresholding in general is slower than OWD and PSS methods. For all the other experiments discussed in this paper, we choose OWD as the optimization method for FHIM. Table 1 and Table 2 shows

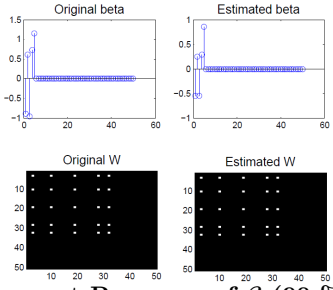

Figure 1: Support Recovery of $\boldsymbol{\beta}$ (90 % sparse) and $\mathbf{W}$ (99 % sparse) for synthetic data Case 1: $n > p$ and $q > n$ where $n = 1000, p = 50, q = 1275$.

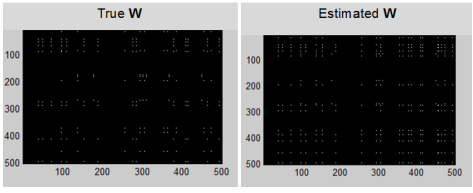

Figure 2: Support Recovery of W (99.5 % sparse) for synthetic data Case 2: $p > n$ where p = 500, n = 100. Online supplementary materials contain high-quality images for this figure.

the performance comparison (in terms of prediction error on test dataset) of FHIM for linear and logistic regression models with respect to the state-of-the-art approaches such as Lasso, Fused Lasso, Trace-Norm and Hierarchical Lasso (HLasso is a general version of SHIM [3]). From tables 1 and 2, we see that FHIM generally outperforms all the state-of-the-art approaches for both linear and logistic pairwise regression models. For $q > n$, we see that test data prediction

	n, p, K	FHIM	Fused Lasso	Lasso	HLasso	Trace norm	Dirty Model
$q > n$	1000, 50, 1	**338.4(14.5)**	425.9(20.7)	474.7(15.3)	354.32 (24.82)	464.4(36.3)	613.5(0.76)
	1000, 50, 5	**343.7(12.9)**	1888.3(121.1)	1922.9(143.9)	889.1 (112.5)	1822.6(99.8)	2453.8(0.76)
	10000, 500, 1	**1093.1(19.5)**	2739.57(155.1)	3896.3(129.5)	-	3887.9(101.1)	4674.7(0.8)
	10000, 500, 5	**1090.76(12.21)**	22720(597.8)	23279.6(231.3)	-	22916.5(321.4)	29214(0.8)
$p > n$	100, 500, 1	**230.49 (50.3)**	1157.2(355.0)	1335.0(159.2)	-	1160.3(299.7)	1651.9(62.6)
	100, 1000, 1	**340.1 (40.02)**	770.9(127.6)	879.1(180.3)	-	699.9(208.7)	808.1(5.1)
	100, 2000, 1	**907.8 (100.1)**	1022.3(406.2)	919.2(132.1)	-	880.42(471.6)	1916.7(63.4)

Table 1: Performance comparison for synthetic data on linear regression model with high-order interactions. Prediction Error (MSE) and Std. deviation of MSE (shown inside brackets) on test data is used to measure the model's performance. For $p >= 500$, Hierarchical Lasso (HLasso) has heavy computational complexity, hence we don't show it's results here.

	n, p, K	FHIM	Fused Lasso	Lasso	HLasso	Trace norm
$q > n$	1000, 50, 1	**0.127 (0.009)**	0.128 (0.017)	0.156 (0.017)	0.136 (0.02)	0.128 (0.016)
	1000, 50, 5	**0.189 (0.03)**	0.227 (0.024)	0.292 (0.042)	0.257 (0.022)	0.503 (0.027)
	10000, 500, 1	**0.135 (0.002)**	0.265 (0.007)	0.161 (0.012)	-	0.225 (0.077)
	10000, 500, 5	**0.390 (0.05)**	0.514 (0.006)	0.507(0.108)	-	0.514 (0.006)
$p > n$	100, 500, 1	**0.325 (0.04)**	0.352 (0.086)	0.4323(0.054)	-	0.40(0.079)
	100, 1000, 1	**0.390 (0.056)**	0.409(0.086)	0.458(0.083)	-	0.438(0.011)

Table 2: Performance comparison for synthetic dataset on logistic regression model with high-order interactions. Misclassification Error on test data is used to measure the model's performance

error for FHIM is consistently lower compared to all other approaches. For $p > n$, FHIM performs slightly better than other approaches, however, the prediction error for all the approaches is high since it's hard to accurately recover the coefficients of main effects and pairwise interactions in very high-dimensional settings.

From figure 1 and table 4, we see that our model performs very well ($F1$ score close to 1) in the support recovery of $\boldsymbol{\beta}$ and $\mathbf{W}$ for the $q > n$ setting. From figure 2 and table 5, we see that our model performs fairly well in the support recovery of $\mathbf{W}$ for $p > n$ setting. We observe that when the tuning parameters are correctly chosen, support recovery of $\mathbf{W}$ works very well when $\mathbf{W}$ is low-rank (see table 4 and 5), and the $F1$ score for the support recovery of $\mathbf{W}$ decreases with increase in rank of $\mathbf{W}$. Table 5 shows that for $q > n$ the greedy strategy of FHIM accurately recovers the rank K of $\mathbf{W}$, while for $p > n$, the greedy strategy might not correctly recover K. This is because the tensor factorization is not unique and slightly correlated variables can enter our model during optimization.

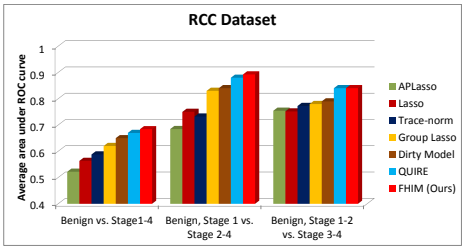

Figure 3: Comparison of the classification performances of different feature selection approaches with our model in identifying the different stages of RCC. We perform five fold cross validation five times and average AUC score is reported.

7.1.5 *Classification Performance on RCC*

In this section, we report systematic experimental results on classification of samples from different stages of RCC. The predictive performance of the markers and pairwise interactions selected by our model (FHIM) is compared against the markers selected by Lasso, All-Pairs Lasso [2], Group Lasso, Dirty model [8] and QUIRE. We use SLEP [13], MALSAR [22] and QUIRE packages for the implementation of these models. The overall performance of the algorithms are shown in Figure 3. In this figure, we report average AUC score for five runs of 5-fold cross validation experiments for cancer stage prediction in RCC. In 5-fold cross validation experiments, we train our model on the four folds to identify the main effects and pairwise interactions and we use the remaining one fold for testing prediction. The average AUC achieved by features selected with our model are 0.68, 0.89 and 0.84 respectively for the three cases discussed in section 7.1.3. We performed pairwise t-tests for the comparisons of our method vs. the other methods, and all p-values are below 0.0075. From figure 3, it is clear that our model outperforms all the other algorithms that do not use prior feature group information for all the three classification cases of RCC prediction. In addition, our model has similar performance to the state-of-the-art technique - QUIRE [14], which uses Gene Ontology based functional annotation for grouping and clustering of genes to identify high order interactions.

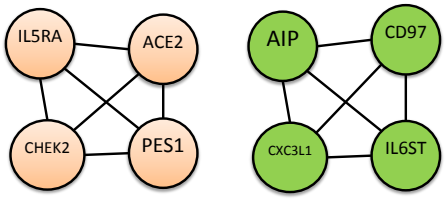

Figure 4: Examples of functional modules for RCC Case 3, induced by markers and interactions discovered by our model and enriched in pathways and functions associated with RCC

7.1.6 *Informative interactions discovered by FHIM*

An investigation of the pairwise interactions identified by our model on RCC dataset reveals that many of these interactions are indeed relevant to the prediction of cancer. Figure 4 shows some of the interactions associated with higher

weighted pairwise co-efficients for Case 3 of RCC classification experiment. The interactions include *CX3CL1- CD97*, *CHEK2-IL5RA* which are known to be related to proteins in blood. *CX3CL1* was recently found to promote breast cancer [17], while *CD97* was found to promote colorectal cancer [19]. We believe these protein interactions might lead to renal cell cancer. Further investigations of the interactions identified by our model might reveal novel protein interactions associated with renal cell cancer and thus leading to testable hypothesis.

7.1.7 Time Complexity

FHIM has $O(np)$ time complexity for algorithm 1. In general, FHIM takes more time than the Lasso approach since we do alternating optimization of $\boldsymbol{\beta}, \boldsymbol{a_k}$. For $q \sim n$ setting with $n = 1000$, $q = 1275$, our OWD learning optimization method on Matlab takes around ~ 1 minute for 5-fold cross-validation, while for $p > n$ with $p = 2000$, $n = 100$, our FHIM model took around 2 hours for 5-fold cross-validation. Our experiments were run on intel i3 dual-core 2.9 GHz CPU with 8 GB RAM.

n, p	OWD	Soft-thres -holding	PSS
100, 500	**230.5**	276.2	239.5
100, 1000	**340.1**	710.5	358.7
100, 2000	**907.8**	1174.1	927.4

Table 3: Comparison of optimization methods for our FHIM model based on test data prediction error

n, p	Sparsity $\boldsymbol{\beta}, \boldsymbol{a}_k$	K	Support recovery $\boldsymbol{\beta}, \boldsymbol{W}$(F1 score)
1000, 50	5, 5	1	1.0, 1.0
1000, 50	5, 5	3	1.0, 0.95
1000, 50	5, 5	5	1.0, 0.82
10000, 500	10, 20	1	0.95, 0.72
10000, 500	10, 20	3	0.80, 0.64
10000, 500	10, 20	5	0.72, 0.55

Table 4: Support recovery of $\boldsymbol{\beta}, \mathbf{W}$

n, p	true K	estimated K	W support recovery F1 score
1000, 50	1	1	1.0
1000, 50	3	3	1.0
1000, 50	5	5	0.8
100, 100	1	2	0.75
100, 500	3	2	0.6
100, 1000	5	4	0.5

Table 5: Recovering K using greedy strategy

8. CONCLUSIONS

In this paper, we proposed a factorization based sparse learning framework called FHIM for identifying high-order feature interactions in linear and logistic regression models, and studied several optimization methods for our model. Empirical experiments on synthetic and real datasets showed that our model outperforms several well-known techniques such as Lasso, Trace-norm, GroupLasso and achieves comparable performance to the current state-of-the-art method - QUIRE, while not assuming any prior knowledge about the data. Our model gives 'interpretable' results for high-order feature interactions on RCC dataset which can be used for biomarker discovery for disease diagnosis.

In the future, we will consider the following directions: (i) We will consider factorization of the weight matrix $\mathbf{W}$ as $\mathbf{W} = \sum_k \boldsymbol{a}_k \boldsymbol{b}_k^T$ and higher-order feature interactions, which is more general, but the optimization problem is non-convex; (ii) We will extend our optimization methods from Single-Task Learning to Multi-Task Learning; (iii) We will consider groupings of features for both Single Task Learning and Multi-Task Learning.

References

[1] G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pages 33–40. ACM, 2007.

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

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

[5] J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

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

[7] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. Pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–332, 2007.

[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] 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.

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

[12] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

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

[14] 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.

[15] M. Schmidt. *Graphical model structure learning with l1-regularization.* PhD thesis, UNIVERSITY OF BRITISH COLUMBIA, 2010.

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

[17] M. Tardaguila, E. Mira, M. A. García-Cabezas, A. M. Feijoo, M. Quintela-Fandino, I. Azcoitia, S. A. Lira, and S. Manes. Cx3cl1 promotes breast cancer via transactivation of the egf pathway. *Cancer research*, 2013.

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

[19] M. Wobus, O. Huber, J. Hamann, and G. Aust. Cd97 overexpression in tumor cells at the invasion front in colorectal cancer (cc) is independently regulated of the canonical wnt pathway. *Molecular carcinogenesis*, 45(11):881–886, 2006.

[20] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714–721, 2009.

[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.

APPENDIX

A. PROOFS FOR SECTION 5.1

PROOF OF LEMMA 5.1:. Let $\eta_n = n^{-1/2} + a_n$ and $\{\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$ be the ball around $\boldsymbol{\theta}^*$, where $\boldsymbol{\delta} = (u_1, ..., u_p, v_{11}, ...v_{Kp})^T = (\mathbf{u}^T, \mathbf{v}^T)^T$. Define

$$D_n(\boldsymbol{\delta}) \equiv Q_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) - Q_n(\boldsymbol{\theta}^*)$$

Where $Q_n(\boldsymbol{\theta}^*)$ is defined in equation (8). For $\boldsymbol{\delta}$ that satisfies $||\boldsymbol{\delta}|| = d$, we have

$$\begin{aligned}
D_n(\boldsymbol{\delta}) =& - L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) \\
&+ n\sum_j \lambda_j^\beta(|\beta_j^* + \eta_n u_j| - |\beta_j^*|) \\
&+ n\sum_{k,l} \lambda_l^{\alpha_k}(|\alpha_{k,l}^* + \eta_n v_{kl})| - |\alpha_{k,l}^*| \\
\geq& - L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) \\
&+ n\sum_{j\in\mathcal{A}_1} \lambda_j^\beta(|\beta_j^* + \eta_n u_j| - |\beta_j^*|) \\
&+ n\sum_{(k,l)\in\mathcal{A}_2} \lambda_l^{\alpha_k}(|\alpha_{k,l}^* + \eta_n v_{kl})| - |\alpha_{k,l}^*| \\
\geq& - L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) \\
&- n\eta_n\sum_{j\in\mathcal{A}_1} \lambda_j^\beta|u_j| - n\eta_n\sum_{(k,l)\in\mathcal{A}_2} \lambda_l^{\alpha_k}|v_{kl}| \\
\geq& - L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) \\
&- n\eta_n^2(\sum_{j\in\mathcal{A}_1}|u_j| + \sum_{(k,l)\in\mathcal{A}_2}|v_{kl}|) \\
\geq& - L_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) + L_n(\boldsymbol{\theta}^*) - n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d \\
=& - [\nabla L_n(\boldsymbol{\theta}^*)]^T(\eta_n\boldsymbol{\delta}) - \frac{1}{2}(\eta_n\boldsymbol{\delta})^T[\nabla^2 L_n(\boldsymbol{\theta}^*)] \\
&(\eta_n\boldsymbol{\delta})(1 + o_p(1)) - n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d
\end{aligned}$$

We used Taylor's expansion in above step. We split the above into three parts and we get:

$$\begin{aligned}
K_1 =& - \eta_n[\nabla L_n(\boldsymbol{\theta}^*)]^T\boldsymbol{\delta} \\
=& - \sqrt{n}\eta_n(\frac{1}{\sqrt{n}}\nabla L_n(\boldsymbol{\theta}^*))^T\boldsymbol{\delta} \\
=& - O_p(n\eta_n^2)\boldsymbol{\delta} \\
K_2 =& \frac{1}{2}n\eta_n^2\{\boldsymbol{\delta}^T[-\frac{1}{n}\nabla^2 L_n(\boldsymbol{\theta}^*)\boldsymbol{\delta}](1 + o_p(1))\} \\
=& \frac{1}{2}n\eta_n^2\{\boldsymbol{\delta}^T[I(\boldsymbol{\theta}^*)\boldsymbol{\delta}](1 + o_p(1))\} \\
K_3 =& - n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d
\end{aligned}$$

Thus,

$$\begin{aligned}
D_n(\boldsymbol{\delta}) \geq& K_1 + K_2 + K_3 \\
=& - O_p(n\eta_n^2)\boldsymbol{\delta} + \frac{1}{2}n\eta_n^2\{\boldsymbol{\delta}^T[I(\boldsymbol{\theta}^*)\boldsymbol{\delta}](1 + o_p(1))\} \\
&- n\eta_n^2(|\mathcal{A}_1| + |\mathcal{A}_2|)d
\end{aligned}$$

We see that K_2 dominates the rest of the terms and is positive since $I(\boldsymbol{\theta})$ is positive definite at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$ from regularity condition (C2). Therefore, for any given $\epsilon > 0$ there exists a large enough constant d such that

$$P\{\inf_{||\delta|| = d} Q_n(\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta}) > Q_n(\boldsymbol{\theta}^*)\} \geq 1 - \epsilon$$

This implies that with probability at-least $1-\epsilon$, there exists a local minimizer in the ball $\{\boldsymbol{\theta}^* + \eta_n\boldsymbol{\delta} : ||\boldsymbol{\delta}|| \leq d\}$. Thus, there exists a local minimizer of $Q_n(\boldsymbol{\theta})$ such that $||\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*|| = O_p(\eta_n)$. □

PROOF OF THEOREM 5.2:. Let us first consider $P(\hat{\alpha}_{\mathcal{A}_2^c} = 0) \to 1$. It is sufficient to show that for any $(k, l) \in \mathcal{A}_2^c$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} < 0 \text{ for } -\epsilon_n < \hat{\alpha}_{k,l} < 0 \tag{11}$$

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} > 0 \text{ for } \epsilon_n > \hat{\alpha}_{k,l} > 0 \tag{12}$$

with probability tending to 1, where $\epsilon_n = Cn^{-1/2}$ and $C > 0$ is any constant. To show (12), notice

$$\begin{aligned}
\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} =& - \frac{L_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} + n\lambda_l^{\alpha_k} sgn(\hat{\alpha}_{k,l}) \\
=& - \frac{L_n(\boldsymbol{\theta}^*)}{\partial\alpha_{k,l}} - \sum_{j=1}^{p(K+1)} \frac{\partial^2 L_n(\boldsymbol{\theta}^*)}{\partial\alpha_{k,l}\partial\theta_j}(\hat{\theta}_j - \theta_j^*)
\end{aligned}$$

$$- \sum_{j=1}^{p(K+1)} \sum_{m=1}^{p(K+1)} \frac{\partial^3 L_n(\tilde{\boldsymbol{\theta}})}{\partial\alpha_{k,l}\partial\theta_j\partial\theta_m}(\hat{\theta}_j - \theta_j^*)(\hat{\theta}_m - \theta_m^*)$$
$$+n\lambda_l^{\alpha_k} sgn(\hat{\alpha}_{k,l})$$

where $\tilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ and $\boldsymbol{\theta}^*$. By regularity conditions (C1)-(C3) and the Lemma 5.1, we have

$$\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} = \sqrt{n}\{O_p(1) + \sqrt{n}\lambda_l^{\alpha_k} sgn(\hat{\alpha}_{k,l})\}$$

As $\sqrt{n}\lambda_l^{\alpha_k} \to \infty$ for $(k,l) \in \mathcal{A}_2^c$ from the assumption, the sign of $\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}}$ is dominated by $sgn(\hat{\alpha}_{k,j})$. Thus,

$$P\left(\frac{\partial Q_n(\hat{\boldsymbol{\theta}}_n)}{\partial\alpha_{k,l}} > 0 \text{ for } 0 < \hat{\alpha}_{k,l} < \epsilon_n\right) \to 1 \text{ as } n \to \infty$$

(11) has identical proof as above. Also, $P(\hat{\boldsymbol{\beta}}_{\mathcal{A}_1^c} = 0) \to 1$ can be proved similarly since in our model $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ are independent of each other. □

PROOF OF THEOREM 5.3. Let $Q_n(\boldsymbol{\theta}_\mathcal{A})$ denote the objective function Q_n only on the $\mathcal{A}-$component of $\boldsymbol{\theta}$, that is $Q_n(\boldsymbol{\theta})$ with $\boldsymbol{\theta}_{\mathcal{A}^c}$. Based on Lemma 5.1 and Theorem 5.2, we have $P(\hat{\boldsymbol{\theta}_{\mathcal{A}^c}} = 0) \to 1$. Thus,

$$P\Big(\arg\min_{\boldsymbol{\theta}_\mathcal{A}} Q_n(\boldsymbol{\theta}_\mathcal{A}) = (\mathcal{A} - \text{component of } \arg\min_{\boldsymbol{\theta}} Q_n(\boldsymbol{\theta}))\Big) \to 1$$

Thus, $\hat{\boldsymbol{\theta}}_\mathcal{A}$ should satisfy

$$\frac{\partial Q_n(\boldsymbol{\theta}_\mathcal{A})}{\partial\theta_j}\Big|_{\boldsymbol{\theta}_\mathcal{A}=\hat{\boldsymbol{\theta}}_\mathcal{A}} = 0 \qquad \forall j \in \mathcal{A} \tag{13}$$

with probability tending to 1. Let $L_n(\boldsymbol{\theta}_\mathcal{A})$ and $P_\lambda(\boldsymbol{\theta}_\mathcal{A})$ denote the log-likelihood function of $\boldsymbol{\theta}_\mathcal{A}$ and the penalty function of $\boldsymbol{\theta}_\mathcal{A}$ respectively so that we have

$$Q_n(\boldsymbol{\theta}_\mathcal{A}) = -L_n(\boldsymbol{\theta}_\mathcal{A}) + nP_\lambda(\boldsymbol{\theta}_\mathcal{A})$$

From (13), we have

$$\nabla_\mathcal{A} Q_n(\hat{\boldsymbol{\theta}}_\mathcal{A}) = -\nabla_\mathcal{A} L_n(\hat{\boldsymbol{\theta}}_\mathcal{A}) + n\nabla_\mathcal{A} P_\lambda(\hat{\boldsymbol{\theta}}_\mathcal{A}) = 0, \tag{14}$$

with probability tending to 1.
Now, consider by Taylor expansion of first term and second terms at $\boldsymbol{\theta}_\mathcal{A} = \boldsymbol{\theta}_\mathcal{A}^*$, we get the following:

$$\begin{aligned}
-\nabla_\mathcal{A} L_n(\hat{\boldsymbol{\theta}}_\mathcal{A}) =& -\nabla_\mathcal{A} L_n(\boldsymbol{\theta}_\mathcal{A}^*) - [\nabla_\mathcal{A}^2 L_n(\boldsymbol{\theta}_\mathcal{A}^*) + o_p(1)] \\
& (\hat{\boldsymbol{\theta}}_\mathcal{A} - \boldsymbol{\theta}_\mathcal{A}^*) \\
=& \sqrt{n}\Bigg[-\frac{1}{\sqrt{n}}\nabla_\mathcal{A} L_n(\boldsymbol{\theta}_\mathcal{A}^*) + \\
& \mathbf{I}(\boldsymbol{\theta}_\mathcal{A}^*)\sqrt{n}(\hat{\boldsymbol{\theta}}_\mathcal{A} - \boldsymbol{\theta}_\mathcal{A}^*) + o_p(1)\Bigg]
\end{aligned}$$

$$\begin{aligned}
n\nabla_\mathcal{A} P_\lambda(\hat{\boldsymbol{\theta}}_\mathcal{A}) =& n\Bigg\{ \begin{bmatrix} \lambda_j^\beta sgn(\beta_j) \\ \lambda_l^{\alpha_k} sgn(\alpha_{k,l}) \end{bmatrix}_{j\in\mathcal{A}_1,(k,l)\in\mathcal{A}_2} \\
& + o_p(1)(\hat{\boldsymbol{\theta}}_\mathcal{A} - \boldsymbol{\theta}_\mathcal{A}^*)\Bigg\} \\
=& \sqrt{n}o_p(1)
\end{aligned}$$

since $\sqrt{n}a_n = o(1)$ and $||\hat{\boldsymbol{\theta}}_\mathcal{A} - \boldsymbol{\theta}_\mathcal{A}^*|| = O_p(n^{-1/2})$

Thus, we get,

$$\begin{aligned}
0 =& \sqrt{n}\Bigg[-\frac{1}{\sqrt{n}}\nabla_\mathcal{A} L_n(\boldsymbol{\theta}_\mathcal{A}^*) \\
& + \mathbf{I}(\boldsymbol{\theta}_\mathcal{A}^*)\sqrt{n}(\hat{\boldsymbol{\theta}}_\mathcal{A} - \boldsymbol{\theta}_\mathcal{A}^*) + o_p(1)\Bigg]
\end{aligned}$$

Therefore, from central limit theorem,

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

□

The proofs for lemma and theorem of section 5.2 are along the same lines as above. Please refer to Appendix section C for more details.

B. COMPUTING ADAPTIVE WEIGHTS

Here, we explain how the adaptive weights $w_j^\beta, w_l^{\alpha_k}$ can be calculated for tuning parameters $\lambda_\beta, \lambda_{\alpha_k}$ in Theoretical properties (Theorems 5.3 & 5.5) of Section 5. Let q be the total number of predictors, let n be total number of instances. When $n > q$, we can compute the adaptive weights $w_j^\beta, w_l^{\alpha_k}$ for tuning parameters $\lambda_j^\beta, \lambda_l^{\alpha_k}$ using ordinary least squares (OLS) estimates of the training observations.

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

where

$$w_j^\beta = |\frac{1}{\hat{\beta}_j^{OLS}}|, \qquad w_l^{\alpha_k} = |\frac{1}{\hat{\alpha_k}_l^{OLS}}|,$$

When $q > n$, the OLS estimates are not available and so we compute the weights using the ridge regression estimates, that is, replacing all the above OLS estimates with the ridge regression estimates. The tuning parameter for ridge regression can be selected using cross-validation. Note, we find $\hat{\alpha_k}_j^{OLS}$ by taking least squares w.r.t to each α_k where $k \in [0, K]$ for some $K \geq trueK$. Without loss of generality we can assume $K = trueK$ for proving the Theoretical properties in section 5. Even if $K \geq trueK$, it does not affect the Theoretical properties since the cardinality of $\mathcal{A}_2(|\mathcal{A}_2|)$ does not affect the root-n consistency (see, proof of lemma 5.1). In practice, K is greedily chosen by algo. 1 in our paper.

C. SUPPLEMENTARY MATERIALS

For interested readers, we provide online supplementary materials at `http://www.cs.toronto.edu/~cuty/FHIM_Supp.pdf` with detailed proofs for Lemma 5.4 and Theorem 5.5. These proofs do not affect the understanding of this paper. We also provide high quality images for figure 1 in the supplementary materials. We provide more details about the experimental settings for state-of-the-art techniques used in this paper.