ORDER-BASED GENERALIZED MULTIVARIATE REGRESSION

Milad Kharratzadeh and Mark Coates

Department of Electrical and Computer Engineering, McGill University

ABSTRACT

In this paper, we consider a generalized multivariate regression problem where the responses are monotonic functions of linear transformations of predictors. We propose a semi-parametric algorithm based on the ordering of the responses which is invariant to the functional form of the transformation function. We prove that our algorithm, which maximizes the rank correlation of responses and linear transformations of predictors, is a consistent estimator of the true coefficient matrix. We also identify the rate of convergence and show that the squared estimation error decays with a rate of $o(1/\sqrt{n})$. We then propose a greedy algorithm to maximize the highly non-smooth objective function of our model and examine its performance through simulations. Finally, we compare our algorithm with traditional multivariate regression algorithms over synthetic and real data.

Index Terms— Generalized Multivariate Regression, Semi-parametric, Kendall's Rank Correlation

1. PROBLEM SETUP

In linear multivariate regression, we have the following model:

$$\mathbf{y}_i^T = \mathbf{x}_i^T\mathbf{B} + \boldsymbol{\epsilon}, \qquad i = 1, \dots, n, \tag{1}$$

where $\mathbf{y}_i \in \mathbb{R}^{q\times 1}$ is the response vector ($q > 1$), $\mathbf{x}_i \in \mathbb{R}^{p\times 1}$ is the predictor vector, $\mathbf{B} \in \mathbb{R}^{p\times q}$ is the coefficient matrix, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^{q\times 1}$ represents the noise with i.i.d. elements that are independent of $\mathbf{x}_i$. In this paper, we consider the following extension of this problem:

$$\mathbf{y}_i^T = U_i(\mathbf{x}_i^T\mathbf{B} + \boldsymbol{\epsilon}_i^T), \qquad i = 1, \dots, n, \tag{2}$$

where U_i: $\mathbb{R} \to \mathbb{R}$ is a non-degenerate monotonic function called the *utility* or *link* function. When the input of U_i is a vector or a matrix, it is implied that U_i is applied separately on each individual element to give the output, which is a vector or matrix of the same size as the input. Without loss of generality, we assume that U_i is an increasing function. We propose a semi-parametric, rank-based approach to estimate $\mathbf{B}$ which is invariant with respect to the functional form of U_i functions. Our approach only uses the ordering of the elements of $\mathbf{y}_i$, which makes it more robust to outliers and heavy-tailed noise compared to traditional regression algorithms. This also makes our approach applicable to cases where the numeric values of $\mathbf{y}_i$ are not available, and only their ordering is known.

We show that it is possible to consistently estimate $\mathbf{B}$ solely based on the ordering of the elements of $\mathbf{y}_i$. Our approach to estimating $\mathbf{B}$ is based on maximizing Kendall's rank correlation of $\mathbf{y}_i^T$ and $\mathbf{x}_i^T\mathbf{B}$. For notational simplicity, we assume that all the link functions are equal and denote them by U; however, all the results presented in this paper hold for the case where there is a separate link function, U_i, for each observation. Let us rewrite (2) in matrix form:

$$\mathbf{Y}_{n\times q} = U(\mathbf{X}_{n\times p}\mathbf{B}_{p\times q} + \mathbf{E}_{n\times q}), \tag{3}$$

where p is the number of predictors, q is the number of responses, and n denotes the number of instances. $\mathbf{x}_i^T$, $\mathbf{y}_i^T$, and $\boldsymbol{\epsilon}_i^T$ correspond, respectively, to the i-th rows of $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{E}$. To find $\mathbf{B}$, we propose to solve:

$$\widehat{\mathbf{B}}_n = \arg\max_{\mathbf{B}} \underbrace{\frac{1}{n\binom{q}{2}} \sum_{i=1}^{n}\sum_{j=1}^{q}\sum_{k=1}^{q} \mathbf{1}(y_{ij}>y_{ik})\mathbf{1}(\mathbf{x}_i^T\mathbf{b}_j>\mathbf{x}_i^T\mathbf{b}_k)}_{S_n(\mathbf{B})}, \tag{4}$$

where $\mathbf{b}_j$ denotes the j-th column of $\mathbf{B}$. The intuition behind this formulation is that since U is increasing and the error is i.i.d. and independent of $\mathbf{x}$, when we have $\mathbf{x}_i^T\mathbf{b}_j > \mathbf{x}_i^T\mathbf{b}_k$, it is more likely to have $y_{ij} > y_{ik}$ than $y_{ij} < y_{ik}$. The term in the summation is zero for $j = k$. Maximizing $S_n(\mathbf{B})$ is equivalent to maximizing the average rank correlation of $\mathbf{y}_i^T$ and $\mathbf{x}_i^T\mathbf{B}$ since $2S_n(\mathbf{B}) - 1$ corresponds to the average over the n observations of the Kendall's rank correlation between $\mathbf{y}_i^T$ and $\mathbf{x}_i^T\mathbf{B}$.

2. MOTIVATING EXAMPLES AND RELATED WORK

2.1. Learning from non-linear measurements

In many practical settings, the measurements or observations are noisy non-linear functions of a linear transformation of an underlying signal. This could be due to the uncertainties and non-linearities of the measurement device or arise from the experimental design. In the statistics and economics literature, this model is known as the single-index model and it has been studied extensively [1–6]. The response in the single-index model is univariate and the form of the link function is sometimes assumed known.

In our model, the response is a vector (multivariate regression) and we assume that the functional form of the link function is unknown. Also, our inference approach only uses the ordering of the elements of the response vector. Recently, it has been shown that under certain assumptions (e.g., when the predictors are drawn from a Gaussian distribution), Lasso with non-linear measurements is equivalent to one with linear measurement with an equivalent input noise proportional to the non-linearity of the link function [7]. Thus, it has been suggested to use Lasso in the non-linear case as if the measurements were linear. Here, and under much more general conditions, we show that our algorithm performs better than a simple application of Lasso to the non-linear problem.

2.2. Learning from the ordering of responses

Our approach is particularly of interest in applications (e.g., surveys) where subjects order a set of items based on their preferences, e.g., people ranking different types of sushi based on their preference [8]. In these scenarios, the underlying model cannot be learned by traditional regression techniques, which require a numerical response. However, our algorithm is directly applicable since it only uses the

978-1-4673-7802-4/16/$31.00 ©2016 IEEE

ordering of the elements of the response vector. Even in the scenarios where the actual values of responses are available (e.g., numerical ratings), it is often more sensible to focus on the ordering rather than striving to learn based on the assigned numerical values. As discussed in [9], there is often no invariant and objective mapping between true preference and observed ratings among subjects, since "each user uses his/her own mapping based on a subjective and variable criterion in his/her own mind". Thus, the mappings might be inconsistent among different subjects. Moreover, the mappings might be inconsistent for a given subject across different items [10]. By using the ordering in training and prediction, we minimize the effects of these inconsistencies.

2.3. Collaborative and content-based filtering

Our work is also related to the problem of personalized recommendation systems, but with important differences. Recommendation systems can be divided into three main categories: content–based filtering, collaborative filtering, and hybrid models [11–13]. Content–based filtering employs the domain knowledge of users and items to predict the ratings. Collaborative filtering does not use any user or item information except a partially observed rating matrix, with rows and vectors corresponding to users and items and matrix elements corresponding to ratings. In general, the rating matrix is extremely sparse, since each user, normally, does not experience and rate all items. Hybrid systems combine collaborative and content–based filtering, e.g., by averaging separate predictions.

If the regression–based framework described in this paper were used in a recommendation system, it would predict each user's ordering of a set of items based on a set of features for that user. These features could include demographic information, user profile data, or ratings of a fixed set of items. Contrary to content–based filtering, our approach does not need domain-specific knowledge about the features of items; this is potentially useful in applications where the items to be ranked are diverse in nature. Also, as opposed to collaborative filtering, we can incorporate user profile data and provide predictions for new users even if they have provided no prior ratings. Thus, our algorithm is different from the problems of collaborative and content-based filtering. It is important to stress that we introduce and study a general semi-parametric multivariate regression method which can be used in recommendation systems, but this is just one of multiple potential applications.

2.4. Maximum rank correlation estimation

In [14], Han considered a problem similar to (2) but with an important difference. His formulation, called Maximum Rank Correlation (MRC) estimation, was stated for the multiple regression setting, where y_i is real-valued rather than a q-vector, and his goal was to maximize the rank correlation across instances. Therefore, the goal in MRC estimation is to capture the ordering of $y_i, i = 1, \ldots, n$ (across instances), whereas in this paper, our goal is to capture the ordering of $y_{ij}, j = 1, \ldots, q$ for a fixed i (for a specific instance, across responses). Considering the ordering across responses enables us to model applications where an instance's ordering (or rating) of a set of items depends exclusively on its predictors. Also, the identifiability and consistency conditions for problem (2) differ significantly from those of the multiple regression problem. There are extensions of the MRC approach (e.g., [15, 16]), but they all are in the multiple regression domain and only differ in how they define the objective function to solve the same problem. Our work differs from them for the same reasons mentioned above.

3. STRONG CONSISTENCY

In this section, we show that the solution of (4) is strongly consistent under certain conditions. $S_n(\mathbf{B})$ is invariant to the multiplication of all elements of $\mathbf{B}$ by a positive constant; i.e., for $c > 0$, $S_n(\mathbf{B}) = S_n(c\mathbf{B})$. The objective function also does not change if the same vector is added to all of the columns of $\mathbf{B}$; i.e., for any $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$, $S_n(\mathbf{B}) = S_n(\mathbf{B} + \boldsymbol{\beta}\mathbb{1}_{1\times q})$, where $\mathbb{1}$ is a vector of all ones. These invariances are expected, since to have a semi-parametric estimate, we target maximizing the rank correlation and ranks are not affected when all the elements are multiplied by a positive constant (c), or are increased/decreased by the same amount ($\mathbf{x}^T\boldsymbol{\beta}$). In other words, since our estimation is semi-parametric in U (and thus, must be invariant to strictly monotonic transformations of observations), $\mathbf{B}$ and $c\mathbf{B} + \boldsymbol{\beta}\mathbb{1}_{1\times q}$ are equivalent. So, we assume that $\|\mathbf{B}\|_F = 1$ (normalization), and that the last column of $\mathbf{B}$ is all zeros (subtracting the last column from all columns). We perform the optimization in (4) over the set

$$\mathcal{B} \triangleq \{\mathbf{B}_{p\times q} : \|\mathbf{B}\|_F = 1 \text{ and } \mathbf{B}_{i,q} = 0 \text{ for } i = 1, 2, \ldots, p\},$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Let us denote the true coefficient matrix by $\mathbf{B}^*$ and without loss of generality assume $\mathbf{B}^* \in \mathcal{B}$; otherwise, we can find $c > 0$ and $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ such that $c\mathbf{B}^* + \boldsymbol{\beta}\mathbb{1}_{1\times q} \in \mathcal{B}$ which gives the same value for objective function as $\mathbf{B}^*$. Also, to have a non-degenerate problem, we assume that $\mathbf{B}^*$ does not have rank 1, since in that case there exist two vectors $\mathbf{u} \in \mathbb{R}^{p\times 1}, \mathbf{v} \in \mathbb{R}^{q\times 1}$ such that $\mathbf{B}^* = \mathbf{u}\mathbf{v}^T$, and $\mathbf{y}^T = U(\mathbf{x}^T\mathbf{u}\mathbf{v}^T + \boldsymbol{\epsilon})$. Therefore, the ordering of the elements of $\mathbf{y}$ will be either the same as the ordering of the elements of $\mathbf{v}$ (if $\mathbf{x}^T\mathbf{u} > 0$) or the reverse of it (if $\mathbf{x}^T\mathbf{u} < 0$) with perturbations due to the noise. So, different observed orderings of the elements of $\mathbf{y}$ are caused merely by noise which is not of interest. Finally, we assume that no two columns of $\mathbf{B}^*$ are equal, because in that case the expected values of the corresponding elements of $\mathbf{y}$ will be the same. Given the model in (3), to prove strong consistency, we need the following three conditions:

- **(C1)** U is a non-degenerate increasing function and changes value at least at one non-zero point (i.e., U is not a step function changing value only at 0).
- **(C2)** The elements of $\mathbf{E}$ are i.i.d. random variables.
- **(C3)** The rows of $\mathbf{X}$ are i.i.d. random p-vectors independent of the elements of $\mathbf{E}$ and have a distribution function F_X such that:
  - **(C3.1)** the support of F_X is not contained in any proper linear subspace of $\mathbb{R}^p$, and
  - **(C3.2)** for all $j \in \{1, 2, \ldots, q\}$ the conditional distribution of x_j given the other components has everywhere positive Lebesgue density.
- **(C4)** $\mathbf{B}^*$ is an interior point of $\mathcal{B}$. Moreover, $\mathbf{B}^*$ has a rank higher than one and no two columns of it are equal.

The second part of condition **(C1)** is needed for the identifiability. However, for all practical purposes, the step function at 0 can be replaced by an approximate function changing value over $[-\epsilon, \epsilon]$ for some $\epsilon > 0$, for which our theoretical results hold. Conditions **(C3.1)** and **(C3.2)** are also required for identifiability, and hold in many settings; e.g., when the rows of $\mathbf{X}$ have a multivariate Gaussian distribution. In **(C4)**, $\mathbf{B}^* \in \mathcal{B}$ implies that its last column is all zeros, and because no two columns are equal, we can conclude that every column except the last one has at least one non-zero element. For some known constant $\eta > 0$ which is less than all the absolute values of these non-zero elements, define

$$\mathcal{B}_\eta \triangleq \{\mathbf{B} : \mathbf{B} \in \mathcal{B}; \forall j \in \{1, \ldots, p\} \;\; \exists i \in \{1, \ldots, p\} \text{ s.t. } |\mathbf{B}_{i,j}| \geq \eta\}.$$

Denoting the solution of (4) over the set $\mathcal{B}_\eta$ by $\widehat{\mathbf{B}}_n$, we prove: $\lim_{n\to\infty} \widehat{\mathbf{B}}_n \to \mathbf{B}^*$. We conduct the proof in three steps. In Lemmas 1 and 2, we prove the identifiability and the convergence of $S_n(\mathbf{B})$ to the expected value of the rank correlation. We then prove the consistency in Theorem 1. The proofs are omitted here but can be found in the extended version of this paper [17].

Lemma 1 (Identifiability)**.** *Given* **(C1)—(C4)**, $\mathbf{B}^*$ *attains the unique maximum of* $E\left[S_n(\mathbf{B}))\right]$ *over the set* $\mathcal{B}$.

Lemma 2 (Convergence)**.** *Given* **(C1)—(C4)***, and denoting*

$$h_i(\mathbf{B}) = \frac{1}{\binom{q}{2}} \sum_{j=1}^{q} \sum_{k=1}^{q} \mathbf{1}(y_{ij} > y_{ik}) \mathbf{1}(\mathbf{x}_i^T \mathbf{b}_j > \mathbf{x}_i^T \mathbf{b}_k),$$

we have: $S_n(\mathbf{B}) \xrightarrow{a.s.} E\left[h_i(\mathbf{B})\right]$.

Theorem 1. *Given* **(C1)—(C4)***, the solution of (4) over the set* $\mathcal{B}_\eta$*,* $\widehat{\mathbf{B}}_n$*, is strongly consistent; i.e.,* $\widehat{\mathbf{B}}_n \to \mathbf{B}^*$ *almost surely*.

Proof. Given the results of Lemmas 1 and 2, we are now ready to prove the consistency of the solution of (4) over $\mathcal{B}_\eta$. We do so by showing that any set, $\mathcal{B}_0 \subset \mathbb{R}^{p\times q}$, that contains $\mathbf{B}^*$, also contains $\widehat{\mathbf{B}}_n$ as $n \to \infty$. Define $\mathcal{B}_1 \triangleq \mathcal{B}_\eta - (\mathcal{B}_0 \cap \mathcal{B}_\eta)$ and $h(\mathbf{B}) \triangleq E\left[h_i(\mathbf{B})\right]$. $\mathcal{B}_1$ is compact and there exists $\zeta = h(\mathbf{B}^*) - \max_{\mathbf{B}\in\mathcal{B}_1} h(\mathbf{B})$ which is always greater than 0 (because $\mathbf{B}^*$ attains the unique maximum and $\mathbf{B}^* \notin \mathcal{B}_1$). From the result of Lemma 2, we know that for any ζ, there is an N, such that for $n > N$, $|S_n(\mathbf{B}) - h(\mathbf{B})| < \zeta/2$ for all $\mathbf{B} \in \mathcal{B}_\eta$ with probability 1. This implies that $\widehat{\mathbf{B}}_n$ cannot be in $\mathcal{B}_1$; because otherwise, we get $h(\mathbf{B}^*) - S_n(\widehat{\mathbf{B}}_n) > \zeta/2$ which is in contradiction with almost sure convergence. Thus, $\widehat{\mathbf{B}}_n \in \mathcal{B}_0$ with probability 1, and since this is true for any $\mathcal{B}_0$, we have $\widehat{\mathbf{B}}_n \to \mathbf{B}^*$ almost surely. □

4. RATE OF CONVERGENCE

For ease of notation, let $\boldsymbol{\theta} \in \mathbb{R}^{p(q-1)}$ be the vectorization of the matrix $\mathbf{B} \in \mathcal{B}$, except the last column which is assumed to be all zero. Thus,

$$\boldsymbol{\theta} \triangleq (B_{1,1}, B_{2,1}, \ldots, B_{p,1}, \ldots, B_{1,q-1}, \ldots, B_{p,q-1}). \quad (5)$$

For $\mathbf{B} \in \mathcal{B}$, the corresponding $\boldsymbol{\theta}$ is in $\boldsymbol{\Theta}$, the set of d-dimensional vectors with norm 1. So, we can denote $\mathbf{B}$ and its columns as functions of $\boldsymbol{\theta}$, and write:

$$h(\mathbf{z}, \boldsymbol{\theta}) = \sum_{j=1}^{q} \sum_{k=1}^{q} \mathbf{1}(y_j > y_k) \mathbf{1}(\mathbf{x}^T \mathbf{b}_j(\boldsymbol{\theta}) > \mathbf{x}^T \mathbf{b}_k(\boldsymbol{\theta})), \quad (6)$$

where $\mathbf{z} = (\mathbf{y}, \mathbf{x}) \in \mathbb{R}^{p+q}$ is the joint vector of predictors and responses for an instance, and $\mathbf{b}_j$ denotes the j'th column of $\mathbf{B}$. Let $\widehat{\boldsymbol{\theta}}_n$ correspond to $\widehat{\mathbf{B}}_n$ and $\boldsymbol{\theta}_0$ correspond to $\mathbf{B}^*$. Based on **(C4)** we know that $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$. In the previous section, we showed that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}_0$. In this section, we study the rate of convergence and show that $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \le o_p(1/\sqrt{n})$, where $\|\cdot\|$ is the Euclidean norm.

The following Lemma plays a critical role in establishing the rate of convergence. Its proof, using results from [18–20], is omitted here due to lack of space, but can be found in [17].

Lemma 3. *For* $\boldsymbol{\theta}$ *in an* $o_p(1)$ *neighborhood of* $\boldsymbol{\theta}_0$*, and* $S(\boldsymbol{\theta}) \triangleq E_{\mathbf{z}}\left[h(\mathbf{z}, \boldsymbol{\theta})\right]$*, we have:*

$$S_n(\boldsymbol{\theta}) = S(\boldsymbol{\theta}) + S_n(\boldsymbol{\theta}_0) - S(\boldsymbol{\theta}_0) + o_p(1/\sqrt{n}). \quad (7)$$

For the next Theorem we require that there exists an $o_p(1)$ neighborhood $\mathcal{A}$ of $\boldsymbol{\theta}_0$ and a constant $\kappa > 0$ for which $S(\boldsymbol{\theta}) - S(\boldsymbol{\theta}_0) \le -\kappa\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ for all $\boldsymbol{\theta} \in \mathcal{A}$. The existence of $\mathcal{A}$ is guaranteed since $\boldsymbol{\theta}_0$ is an interior point of $\boldsymbol{\Theta}$. Also, $\kappa > 0$ exists if:

(C5) The matrix $\nabla_2 S(\boldsymbol{\theta}_0)$ is negative definite.

Theorem 2. *Assume that there exists an* $o_p(1)$ *neighborhood* $\mathcal{A}$ *of* $\boldsymbol{\theta}_0$ *and a constant* $\kappa > 0$ *for which* $S(\boldsymbol{\theta}) - S(\boldsymbol{\theta}_0) \le -\kappa\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|^2$ *for all* $\boldsymbol{\theta} \in \mathcal{A}$*. Then, the squared estimation error decays with a rate faster than* $1/\sqrt{n}$*:* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \le o_p(1/\sqrt{n})$.

Proof. By definition of $\widehat{\boldsymbol{\theta}}_n$ we have: $0 \le S_n(\widehat{\boldsymbol{\theta}}_n) - S_n(\boldsymbol{\theta}_0)$. Rewriting this inequality using the result of Lemma 3 gives:

$$0 \le S(\widehat{\boldsymbol{\theta}}_n) - S(\boldsymbol{\theta}_0) + o_p(1/\sqrt{n}) \le -\kappa\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 + o_p(1/\sqrt{n}). \quad (8)$$

which gives us $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \le o_p(1/\sqrt{n})$. □

Corollary 1. *Given* **(C1)–(C5)***, we have* $\|\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \le o_p(1/\sqrt{n})$.

5. OPTIMIZATION

We have shown that solving (4) provides a consistent estimate of $\mathbf{B}^*$. However, the objective function is very non-smooth (the sum of many step functions) and finding $\widehat{\mathbf{B}}_n$ can be challenging. In this section, we propose a fast, greedy algorithm to solve (4). First, consider the following maximization problem:

$$\widehat{x} = \arg\max_{x\in\mathbb{R}} \quad \sum_{t=1}^{T} \mathbf{1}(u_t + v_t x > 0), \quad (9)$$

where $\{u_t\}_{t=1}^T$ and $\{v_t\}_{t=1}^T$ are given real numbers. We propose an $O(T \log T)$ algorithm to find $\widehat{x}$. The algorithm works as follows. First, we sort the sequence $\{-u_t/v_t\}_{t=1}^T$, in $O(T \log T)$. Then, we start x from a value less than the first sorted point, i.e., $\min\{-u_t/v_t; t = 1, \ldots, T\}$, and at each step, move forward to the next smallest point. At each step, we cumulatively add or subtract 1 depending on the sign of v_t (i.e., depending on whether one of the step functions went from 0 to 1 or vice-versa) and keep track of the largest cumulative sum seen so far, and the value of x for which that maximum happened. After going through all T points, the largest observed cumulative sum is equal to the maximum value of the objective function, and the corresponding value of x is $\widehat{x}$. Since the objective function is piecewise constant, its maximum is attained over an interval; we set $\widehat{x}$ to the center of this interval. Therefore, we can solve (4) in $O(T \log T)$.

We use an alternating maximization scheme to solve (4). We initialize the elements of $\mathbf{B}$ randomly (drawn from i.i.d. normal) and go through all its elements one-by-one and update them to maximize $S_n(\mathbf{B})$ while the other elements are kept fixed. We can show that each of these optimization problems are of the form (9) and can be solved easily. Our algorithm is greedy and it may converge to a local maximum. We can alleviate this problem to some degree by starting the algorithm from different random initial points and choosing the best result. In the next section, we show that our proposed alternating maximization scheme is successful in providing better predictive results for both synthetic and real data.

Improvement over		LS	LTS [21]	LASSO [22]	SRRR [23]	ElasticNet [24]	Ridge
E1	median	0.11	0.18	0.09	0.57	0.52	0.50
	95% CI	[0.02, 0.30]	[0.09, 0.51]	[0.01, 0.26]	[0.16, 1.2]	[0.27, 0.86]	[0.24, 0.93]
E2	median	0.11	0.16	0.09	0.62	0.54	0.51
	95% CI	[0.05, 0.41]	[0.07, 0.37]	[0.04, 0.34]	[0.24, 1.2]	[0.27, 0.90]	[0.24,0.96]
E3	median	0.11	0.17	0.08	0.59	0.48	0.50
	95% CI	[0.01, 0.32]	[0.07, 0.44]	[0.00, 0.29]	[0.21, 1.22]	[0.23, 0.93]	[0.19, 0.88]

Table 1: Improvements in test rank correlation achieved by our algorithm over state-of-the-art multivariate regression techniques. We report the statistics of $c_1 - c_2$, where c_1 and c_2 are respectively the test rank correlations of our algorithm and the correlation of the other algorithm.

6. EXPERIMENTS

6.1. Synthetic data

Compared with traditional linear multivariate regression algorithms, our method has two important differences. First, the objective function is defined as the Kendall's rank correlation between responses and their estimates rather than the sum-of-squares error. The second difference is in the assumed underlying model. We consider a model where responses are related to the predictors through an unknown, potentially non-linear function, U, whereas in the traditional techniques, the relationship is assumed to be linear (or in the generalized regression setting, a known function, e.g., $\log$, of inputs).

In this section, we compare our algorithm with some state-of-the-art multivariate regression techniques. We assume a model as in (3) with 10 predictors and 10 responses. The elements of the predictor matrix are sampled from a uniform distribution between 0 and 100. Matrix $\mathbf{B}$ is sparse with density 60% and its elements are drawn from a standard uniform distribution. The utility function is a sigmoid: $U(x) = 1/(1 + e^{-x})$. We consider three noise settings where the elements of $\mathbf{E}$ are drawn independently from a:

- **(E1)** Gaussian distribution $\mathcal{N}(0, 1)$.
- **(E2)** t-student distribution with $\nu = 1$.
- **(E3)** Gaussian mixture model; with probability 0.8 from a $\mathcal{N}(0, .2)$ and with probability $.2$ from a $\mathcal{N}(1, .2)$.

In order to have a consistent signal-to-noise ratio in various scenarios, the elements of $\mathbf{E}$ are scaled such that the norm of the noise matrix is 20% of the norm of the signal matrix. **E1** corresponds to a general setting where the noise is Gaussian; **E2** simulates a heavy-tailed noise which is present in many practical settings; and **E3** simulates a case where 20% of the data points are corrupted with larger noise (i.e., simulating outliers).

The learning is done over 30 instances and the test is done on a separate set of 20 instances. We use 10 random initializations for our algorithm. The algorithms against which we compare our method are Least Squares (LS), a robust version of LS called Least Trimmed Squares (LTS) [21], LASSO [22], Sparse Reduced Rank Regression (SRRR) [23], elastic net [24], and regressions with ridge regularization. The regularization parameters are achieved via 5–fold cross-validation. We run each experiment 100 times and report the median and 95% confidence intervals of the improvements in the test rank correlation in Table 1. In essence, we run a paired hypothesis test comparing our algorithm against each of the algorithms in Table 1, and report the median, 2.5'th percentile, and 97.5'th percentile of the test statistic, $c_1 - c_2$, where c_1 and c_2 are respectively the test rank correlations of our algorithm and of the other algorithm. We observe that in all cases, our algorithm performs better than other algorithms.

6.2. Real data

We consider the sushi preference dataset [25]. This dataset includes the preference ordering of 10 sushi types by 5000 subjects and the demographic information about these subjects. For each subject, the predictors are: gender (male/female), age group (6 in total), region lived in for the longest period until 15 years old (11 in total), and region currently living (11 in total). Features are represented with a binary indicator vector, e.g., for gender, we use $(0, 1)$ for males and $(1, 0)$ for females. The goal of our prediction task is to estimate the ordering of the 10 sushi types for a new subject only based on their demographic information. From the 5000 users, we choose 2500 of them at random as the training set and keep the the rest as the test set. We take the average Kendall correlation between the rows of predicted and true orderings for the test set as the performance metric. We repeat this random division 100 times to achieve bootstrapped confidence intervals for the performance metric.

In order to compare our algorithm to other regression–based algorithms, we need to transform the ordering into ratings. We use the technique described in [9] and assign the ratings $1/11, \ldots, 10/11$ to the least to most preferred items. Then, for the algorithms that performed well in the simulation study (see Table 1), we follow the same training and testing procedure as explained above. We also compare the results to a K nearest neighbor technique (KNN) where the feature vector of a new user is compared with the available users to identify the K most similar users (in terms of the Euclidean distance), and then its ratings are calculated by averaging the ratings of those K neighbors. Regularization parameters are found via 5–fold cross–validation. The results are shown in Table 2. As we observe, our algorithm outperforms other with high statistical significance.

	Order-based	LASSO	LS	SRRR	KNN
Median	0.34	0.31	0.31	0.18	0.31
95% CI	[0.33, 0.35]	[0.30, 0.32]	[0.30, 0.32]	[0.02, 0.27]	[0.31, 0.32]

Table 2: Comparison of performance for the Sushi dataset.

7. CONCLUSION

In this paper, we considered a generalized regression problem where the responses are monotonic functions of a linear transformation of the predictors. We proposed a semi-parametric method based on rank correlation, which is invariant with respect to the functional form of the underlying monotonic function, to estimate the linear transformation. We showed that the solution to our formulated problem is a consistent estimator of the true matrix and identified the convergence rate. To find the solution, we need to maximize a highly non-smooth function. We proposed a greedy algorithm to solve that problem, and showed its success over simulated and real data.

8. REFERENCES

[1] Ker-Chau Li and Naihua Duan, "Regression analysis under link violation," *The Annals of Statistics*, vol. 17, pp. 1009–1052, 1989.

[2] Hidehiko Ichimura, "Semiparametric least squares (sls) and weighted sls estimation of single-index models," *Journal of Econometrics*, vol. 58, no. 1, pp. 71–120, 1993.

[3] Yingcun Xia and Wolfgang Härdle, "Semi-parametric estimation of partially linear single-index models," *Journal of Multivariate Analysis*, vol. 97, no. 5, pp. 1162–1184, 2006.

[4] Michel Delecroix, Marian Hristache, and Valentin Patilea, "On semiparametric m-estimation in single-index regression," *Journal of Statistical Planning and Inference*, vol. 136, no. 3, pp. 730–769, 2006.

[5] Xinyang Yi, Zhaoran Wang, Constantine Caramanis, and Han Liu, "Optimal linear estimation under unknown nonlinear transform," *arXiv preprint arXiv:1505.03257*, 2015.

[6] Peter Radchenko, "High dimensional single index models," *Journal of Multivariate Analysis*, vol. 139, pp. 266–282, 2015.

[7] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi, "Lasso with non-linear measurements is equivalent to one with linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 3402–3410.

[8] Toshihiro Kamishima, Hideto Kazawa, and Shotaro Akaho, "A survey and empirical comparison of object ranking methods," in *Preference learning*, pp. 181–201. Springer, 2011.

[9] Toshihiro Kamishima and Shotaro Akaho, "Nantonac collaborative filtering: A model-based approach," in *Proc. ACM Conf. on Recommender Systems*, 2010, pp. 273–276.

[10] Oscar Luaces, Gustavo F Bayón, José R Quevedo, Jorge Díez, Juan José Del Coz, and Antonio Bahamonde, "Analyzing sensory data using non-linear preference learning with feature subset selection," in *Proc. European Conf. Machine Learning*, 2004, pp. 286–297.

[11] Pasquale Lops, Marco De Gemmis, and Giovanni Semeraro, "Content-based recommender systems: State of the art and trends," in *Recommender Systems Handbook*, pp. 73–105. Springer, 2011.

[12] Joonseok Lee, Mingxuan Sun, and Guy Lebanon, "A comparative study of collaborative filtering algorithms," *arXiv preprint arXiv:1205.3193*, 2012.

[13] Jesús Bobadilla, Fernando Ortega, Antonio Hernando, and Abraham Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[14] Aaron K Han, "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, vol. 35, no. 2, pp. 303–316, 1987.

[15] Christopher Cavanagh and Robert P Sherman, "Rank estimators for monotonic index models," *Journal of Econometrics*, vol. 84, no. 2, pp. 351–381, 1998.

[16] Jason Abrevaya, "Pairwise-difference rank estimation of the transformation model," *Journal of Business & Economic Statistics*, vol. 21, no. 3, pp. 437–447, 2003.

[17] Milad Kharratzadeh and Mark Coates, "Semi-parametric order-based generalized multivariate regression," 2016, Technical Report, McGill University, available at http://networks.ece.mcgill.ca/pubs.

[18] Ariel Pakes and David Pollard, "Simulation and the asymptotics of optimization estimators," *Econometrica*, vol. 57, pp. 1027–1057, 1989.

[19] Deborah Nolan and David Pollard, "U-processes: rates of convergence," *The Annals of Statistics*, vol. 15, pp. 780–799, 1987.

[20] Robert P Sherman, "Maximal inequalities for degenerate u-processes with applications to optimization estimators," *The Annals of Statistics*, vol. 22, pp. 439–459, 1994.

[21] Pavel Čížek and Jan Ámos Víšek, "Least trimmed squares," in *XploRe – Application Guide*, pp. 49–63. Springer Berlin Heidelberg, 2000.

[22] Robert J. Tibshirani, "Regression shrinkage and selection via the lasso: a retrospective," *Journal of the Royal Statistical Society: Series B*, vol. 73, no. 3, pp. 273–282, 2011.

[23] Lisha Chen and Jianhua Z Huang, "Sparse reduced-rank regression for simultaneous dimension reduction and variable selection," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1533–1545, 2012.

[24] Hui Zou and Trevor Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 2, pp. 301–320, 2005.

[25] Sushi Preference Data Sets, *http://www.kamishima.net/sushi/*.