Sparse High-Dimensional Isotonic Regression

David Gamarnik and Julia Gaudio

Introduction

Definition 1 *Let $A \subseteq [d]$. Let $x, y \in \mathbb{R}^d$. Write $x \preceq_A y$ if $x_i \leq y_i$ for all $i \in A$.*

Definition 2 *Let $s < d$. We say that a function $f : \mathbb{R}^d \to \mathbb{R}$ is s-*sparse coordinate-wise monotone *if there exists a set A with cardinality s, such that $x \preceq_A y \implies f(x) \leq f(y)$ for all $x, y \in \mathbb{R}^d$. We call A the* set of active coordinates.

The *sparse isotonic regression problem* is to estimate the function f, knowing the sparsity level s but not the set A. We assume throughout that the function $f(x)$ is not constant with respect to any of the active coordinates.

Noise Models

We consider two noise models. In either model, the input random variable X is uniform on $\mathcal{X} = [0,1]^d$. The noise W is zero-mean and independent from X.
Noisy Output Model: $Y = f(X) + W$. Let $\mathcal{R}$ be the range of f and let $\text{supp}(W)$ be the support of W. We assume that both $\mathcal{R}$ and $\text{supp}(W)$ are bounded. Without loss of generality, let $\mathcal{R} + \text{supp}(W) \subseteq [0,1]$.
Noisy Input Model: $Y = f(X + W)$. We additionally assume that the coordinates of W are independent. We exclusively consider the classification problem, namely $f : \mathbb{R}^d \to \{0,1\}$.
In either noise model, we assume that n independent samples $(X_1, Y_1), \ldots, (X_n, Y_n)$ are given.

Contributions

We introduce algorithms to obtain estimators $\hat{f}_n$ for f.
Simultaneous Algorithm: Determines the active coordinates and the estimated function values in a single optimization formulation based on integer programming.
Two-Stage Algorithm: Determines the active coordinates via a linear program, and then uses these coordinates to separately estimate the function values.
We bound the tail probability $\mathbb{P}\left(\|\hat{f}_n - f\|_2 > \epsilon\right)$.

Simultaneous Algorithm

The simultaneous algorithm solves the following problem.

$$\min_{A,F} \sum_{i=1}^{n} (Y_i - F_i)^2 \tag{1}$$
$$\text{s.t. } |A| = s \tag{2}$$
$$F_i \leq F_j \qquad \text{if } X_i \preceq_A X_j \tag{3}$$
$$F_i \in \mathcal{R} \qquad \forall i \tag{4}$$

The estimated function $\hat{f}_n$ is determined by interpolating from the pairs $(X_1, F_1), \ldots, (X_n, F_n)$, setting

$$\hat{f}_n(x) = \max\{F_i : X_i \preceq x\}.$$

In other words, we identify all points X_i such that $X_i \preceq x$ and select the smallest consistent function value.

Definition 3 *For inputs $X_1, \ldots, X_n$, let $q(i,j,k) = 1$ if $X_{i,k} > X_{j,k}$, and $q(i,j,k) = 0$ otherwise.*

Problem (1)-(4) can be encoded as a single mixed-integer convex minimization. Binary variables v_k indicate the estimated active coordinates. The variables F_i represent the estimated function values at data points X_i.

Algorithm 1 Integer Programming Isotonic Regression (IPIR)

1: Solve the following optimization problem.

$$\min_{v,F} \sum_{i=1}^{n} (Y_i - F_i)^2$$
$$\sum_{k=1}^{d} v_k = s$$
$$\sum_{k=1}^{d} q(i,j,k) v_k \geq F_i - F_j \qquad \forall i,j \in \{1,\ldots,n\}$$
$$v_k \in \{0,1\} \qquad \forall k \in \{1,\ldots,d\}$$
$$F_i \in \mathcal{R} \qquad \forall i \in \{1,\ldots,n\}$$

2: Return the function $\hat{f}_n(x) = \max\{F_i : X_i \preceq x\}$.

Two-Stage Algorithm

Algorithm 2 determines the active coordinates one at a time. Once a coordinate i is included in the set of active coordinates, variable v_i is set to zero in future iterations.

Algorithm 2 Sequential Linear Programming Support Recovery (S-LPSR)

1: $B \leftarrow \emptyset$
2: **while** $|B| < s$ **do**
3: Solve the following optimization problem:

$$\min \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{d} c_k^{ij}$$
$$\sum_{k=1}^{d} v_k = 1$$
$$v_i = 0 \qquad \forall i \in B$$
$$\sum_{k=1}^{d} q(i,j,k) \left(v_k + c_k^{ij}\right) \geq 1$$
$$\text{if } Y_i > Y_j, \sum_{k=1}^{d} q(i,j,k) \geq 1$$
$$0 \leq v_k \leq 1 \qquad \forall k \in \{1,\ldots,d\}$$
$$c_k^{ij} \geq 0 \qquad \forall i,j \in [n], k \in [d]$$

4: Identify i^* such that $v_{i^*} = \max_i\{v_i\}$, breaking ties arbitrarily. Set $B \leftarrow B \cup \{i_{\max}\}$.
5: **end while**
6: Return $\hat{A} = B$.

Algorithm 3 Two Stage Isotonic Regression (TSIR)

1: Estimate $\hat{A}$ by using Algorithm 2. Let $v_k = 1$ if $k \in \hat{A}$ and $v_k = 0$ otherwise.
2: Solve the following optimization problem.

$$\min \sum_{i=1}^{n} (Y_i - F_i)^2$$
$$\text{s.t. } \sum_{k=1}^{d} q(i,j,k) v_k \geq F_i - F_j \qquad \forall i,j \in \{1,\ldots,n\}$$
$$F_i \in \mathcal{R} \qquad \forall i \in \{1,\ldots,n\}$$

3: Return the function $\hat{f}_n(x) = \max\{F_i : X_i \preceq x\}$.

Statistical Guarantees

We state the results for the Noisy Output Model.

Theorem 1 *The L_2 error of the estimator $\hat{f}_n$ obtained from Algorithm 1 is upper bounded as*

$$\mathbb{P}\left(\|\hat{f}_n - f\|_2 \geq \epsilon\right) \leq 6\binom{d}{s} \exp\left[\left(\frac{128\log(2)}{\epsilon^2} + 2^{\frac{64}{\epsilon^2}} 2^s\right) n^{\frac{s-1}{s}} - \frac{\epsilon^4 n}{512}\right].$$

Corollary 1 *When $n = \max\left\{e^{\omega(s^2)}, \omega(s\log(d))\right\}$, the estimator $\hat{f}_n$ from Algorithm 1 is consistent. In particular, if the sparsity level s is constant, the sample complexity is only logarithmic in the dimension.*

Definition 4 *Let $Y_1 = f(X_1) + W_1$ and $Y_2 = f(X_2) + W_2$ be two independent samples from the model. For $k \in A$, let*

$$p_k \triangleq \mathbb{P}\left(Y_1 > Y_2 \mid q(1,2,k) = 1\right) - \mathbb{P}\left(Y_1 < Y_2 \mid q(1,2,k) = 1\right).$$

Assume without loss of generality that $A = \{1,2,\ldots,s\}$ and $p_1 \leq p_2 \leq \cdots \leq p_s$.

Theorem 2 *Let B be the set of indices corresponding to running Algorithm 2' using n samples. Then it holds that $B = A$ with probability at least*

$$1 - ds\exp\left(-\frac{p_1^2 n}{64 s^3}\right).$$

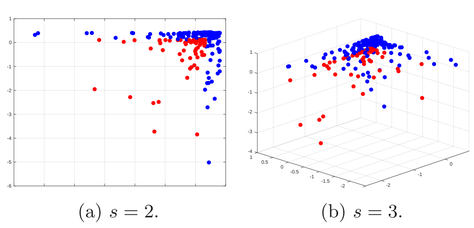
Corollary 2 *Assume that $p_1 = \Theta(1)$. Consider running Algorithm 3 using n samples for sequential recovery. Let $m = \frac{n}{s}$. Consider using an additional m samples for function value estimation. If $n = \max\left\{\omega(s^3\log(d)), se^{\omega(s^2)}\right\}$, then $\hat{f}_{n+m}$ is a consistent estimator.*

Experiments

The presence or absence of a disease is believed to follow a monotone relationship with respect to gene expression. In order to assess the applicability of our sparse monotone regression approach, we apply it to cancer classification using gene expression data. The motivation for using a *sparse* model for disease classification is that certain genes should be more responsible for disease than others.

Table 1: Comparison of classifier success rates on COSMIC data. Top row data is according to the "min" interpolation rule and bottom row data is according to the "max" interpolation rule.

n	IPIR $s=$ 1	2	3	4	5	TSIR + S-LPSR $s=$ 1	2	3	4	5
100	83.1 83.9	84.6 91.8	76.8 91.0	66.2 85.7	53.8 75.7	82.4 82.9	84.6 90.4	77.8 88.9	73.0 87.4	65.4 83.3
200	85.4 85.8	88.1 92.6	84.3 96.4	73.9 88.9	62.7 83.9	85.4 85.8	89.3 94.5	86.7 95.9	81.2 95.3	76.9 93.0
300	- -	- -	- -	- -	- -	84.7 85.1	91.7 94.2	89.0 95.6	84.4 95.9	80.2 94.8
400	- -	- -	- -	- -	- -	85.6 85.8	91.8 94.0	89.7 95.7	87.3 96.4	81.7 95.7



(a) $s = 2$. (b) $s = 3$.

Figure 1: Illustration of the TSIR + S-LPSR algorithm. Blue and red markers correspond to lung and skin cancer, respectively.

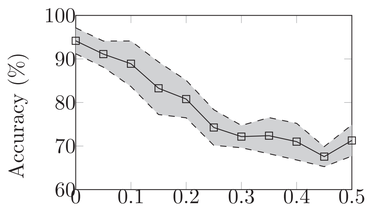


Figure 2: Robustness to error of TSIR + S-LPSR.