Sparse Shrunk Additive Models

Guodong Liu(University of Pittsburgh), Hong Chen (Huazhong Agricultural Univerisity), Heng Huang (University of Pittsburgh)

June 14, 2020

1. Motivation

Deep models have made great progress in learning large dataset, however, statistical models could do better in smaller ones. Also, statistical models usually show better interpretability.

1. Motivation

Deep models have made great progress in learning large dataset, however, statistical models could do better in smaller ones. Also, statistical models usually show better interpretability.

- Linear model.
  - Linear assumption is too restricted.
  - The non-linear fact in applications.

1. Motivation

Deep models have made great progress in learning large dataset, however, statistical models could do better in smaller ones. Also, statistical models usually show better interpretability.

- Linear model.
  - Linear assumption is too restricted.
  - The non-linear fact in applications.
- Generalized additive model.
  - Nonparametric extensions of linear models.
  - Flexible and adaptive to high dimensional data.

1. Motivation

Deep models have made great progress in learning large dataset, however, statistical models could do better in smaller ones. Also, statistical models usually show better interpretability.

- Linear model.
  - Linear assumption is too restricted.
  - The non-linear fact in applications.
- Generalized additive model.
  - Nonparametric extensions of linear models.
  - Flexible and adaptive to high dimensional data.
  - Univariate smooth component function.
  - Pre-defined group structure information.

2. Contribution

- Propose a uniform framework to bridge sparse feature selection, sparse sample selection, and feature interaction structure learning tasks.

2. Contribution

- Propose a uniform framework to bridge sparse feature selection, sparse sample selection, and feature interaction structure learning tasks.
- Provided Generalization bound on the excess risk under mild conditions, which implies the fast convergence rate can be achieved.

2. Contribution

- Propose a uniform framework to bridge sparse feature selection, sparse sample selection, and feature interaction structure learning tasks.
- Provided Generalization bound on the excess risk under mild conditions, which implies the fast convergence rate can be achieved.
- Derived the necessary and sufficient condition to characterize the sparsity of SSAM.

3. Algorithm: Sparse Shrunk Additive Models

- Let $\mathcal{X} \subset \mathbb{R}^n$ be a explanatory feature space and let $\mathcal{Y} \subset [-1, 1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.

3. Algorithm: Sparse Shrunk Additive Models

- Let $\mathcal{X} \subset \mathbb{R}^n$ be a explanatory feature space and let $\mathcal{Y} \subset [-1, 1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- For any given $1 \leq k \leq n$ and $\{1, 2, ..., n\}$, we denote $d = \binom{n}{k}$ as the number of index subset with k elements. Let $x^{(j)} \in \mathbb{R}^k$ be a subset of x with k features and denote its corresponding space as $\mathcal{X}^{(j)}$.

3. Algorithm: Sparse Shrunk Additive Models

- Let $\mathcal{X} \subset \mathbb{R}^n$ be a explanatory feature space and let $\mathcal{Y} \subset [-1, 1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- For any given $1 \leq k \leq n$ and $\{1, 2, ..., n\}$, we denote $d = \binom{n}{k}$ as the number of index subset with k elements. Let $x^{(j)} \in \mathbb{R}^k$ be a subset of x with k features and denote its corresponding space as $\mathcal{X}^{(j)}$.
- Let $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ be a continuous function satisfying $\|K^{(j)}\|_\infty < +\infty$.

3. Algorithm: Sparse Shrunk Additive Models

- Let $\mathcal{X} \subset \mathbb{R}^n$ be a explanatory feature space and let $\mathcal{Y} \subset [-1, 1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- For any given $1 \leq k \leq n$ and $\{1, 2, ..., n\}$, we denote $d = \binom{n}{k}$ as the number of index subset with k elements. Let $x^{(j)} \in \mathbb{R}^k$ be a subset of x with k features and denote its corresponding space as $\mathcal{X}^{(j)}$.
- Let $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ be a continuous function satisfying $\|K^{(j)}\|_\infty < +\infty$.
- For any given $\mathbf{z}$, we define the data dependent hypothesis space as: $\mathcal{H}_{\mathbf{z}} = \{f : f(x) = \sum_{j=1}^d f^{(j)}(x^{(j)}), f^{(j)} \in \mathcal{H}_{\mathbf{z}}^{(j)}\}$, where $\mathcal{H}_{\mathbf{z}}^{(j)} = \{f^{(j)} = \sum_{i=1}^m \alpha_i^{(j)} K^{(j)}(x_i^{(j)}, \cdot) : \alpha_i^{(j)} \in \mathbb{R}\}$

3. Algorithm: Sparse Shrunk Additive Models

- Let $\mathcal{X} \subset \mathbb{R}^n$ be a explanatory feature space and let $\mathcal{Y} \subset [-1, 1]$ be the response set. Let $\mathbf{z} := \{z_i\}_{i=1}^m = \{(x_i, y_i)\}_{i=1}^m$ be independent copies of a random sample (x, y) following an unknown intrinsic distribution ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- For any given $1 \leq k \leq n$ and $\{1, 2, ..., n\}$, we denote $d = \binom{n}{k}$ as the number of index subset with k elements. Let $x^{(j)} \in \mathbb{R}^k$ be a subset of x with k features and denote its corresponding space as $\mathcal{X}^{(j)}$.
- Let $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ be a continuous function satisfying $\|K^{(j)}\|_\infty < +\infty$.
- For any given $\mathbf{z}$, we define the data dependent hypothesis space as: $\mathcal{H}_\mathbf{z} = \{f : f(x) = \sum_{j=1}^d f^{(j)}(x^{(j)}), f^{(j)} \in \mathcal{H}_\mathbf{z}^{(j)}\}$, where $\mathcal{H}_\mathbf{z}^{(j)} = \{f^{(j)} = \sum_{i=1}^m \alpha_i^{(j)} K^{(j)}(x_i^{(j)}, \cdot) : \alpha_i^{(j)} \in \mathbb{R}\}$
- Denote $\|f^{(j)}\|_{\ell_1} = \inf \left\{ \sum_{t=1}^m |\alpha_t^{(j)}| : f^{(j)} = \sum_{t=1}^m \alpha_t^{(j)} K^{(j)}(x_t^{(j)}, \cdot) \right\}$, and $\|f\|_{\ell_1} := \sum_{j=1}^d \|f^{(j)}\|_{\ell_1}$ for $f = \sum_{j=1}^d f^{(j)}$.

3. Algorithm: Sparse Shrunk Additive Models

Predictor of SSAM

$$f_{\mathbf{z}} = \sum_{j=1}^{d} f_{\mathbf{z}}^{(j)} = \sum_{j=1}^{d}\sum_{t=1}^{m} \hat{\alpha}_t^{(j)} K^{(j)}(x_t^{(j)}, \cdot)$$

where, for $1 \le t \le m$ and $1 \le j \le d$,

$$\begin{aligned} \{\hat{\alpha}_t^{(j)}\} = \arg\min_{\alpha_t^{(j)} \in \mathbb{R}, t, j} \Big\{ \lambda \sum_{j=1}^{d}\sum_{t=1}^{m} |\alpha_t^{(j)}| \\ + \frac{1}{m}\sum_{i=1}^{m}\Big(y_i - \sum_{j=1}^{d}\sum_{t=1}^{m} \alpha_t^{(j)} K^{(j)}(x_t^{(j)}, x_i^{(j)})\Big)^2 + \Big\}. \end{aligned} \tag{1}$$

3. Algorithm: Sparse Shrunk Additive Models

SSAM from the viewpoint of function approximation

$$f_{\mathbf{z}} = \arg\min_{f \in \mathcal{H}_{\mathbf{z}}} \left\{ \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2 + \lambda \|f\|_{\ell_1} \right\}.$$

4. Theoretical Analysis: Assumptions

Assumption 1:

Assume that $f_\rho = \sum_{j=1}^{d} f_\rho^{(j)}$, where for each $j \in \{1, 2, ..., d\}$, $f_\rho^{(j)} : \mathcal{X}^{(j)} \to \mathbb{R}$ is a function of the form $f_\rho^{(j)} = L^r_{\tilde{K}^{(j)}}(g_\rho^{(j)})$ with some $r > 0$ and $g_\rho^{(j)} \in L^2_{\rho_{\mathcal{X}^{(j)}}}$.

Assumption 2:

For each $j \in \{1, 2, ..., d\}$, the kernel function $K^{(j)} : \mathcal{X}^{(j)} \times \mathcal{X}^{(j)} \to \mathbb{R}$ is $\mathcal{C}^s$ with some $s > 0$ satisfying:

$$\|K^{(j)}(u, v) - K^{(j)}(u, v')\| \le c_s \|v - v'\|_2^s, \forall u, v, v' \in \mathcal{X}^{(j)}$$

for some positive constant c_s.

4. Theoretical Analysis: Theorems

Theorem 1

Let Assumptions 1 and 2 be true. For any $0 < \delta < 1$, with confidence $1-\delta$, there exists positive constant $\tilde{c}_1$ independent of m, δ such that:

(1) If $r \in (0, \frac{1}{2})$ in Assumption 1, setting $\lambda = m^{-\theta_1}$ with $\theta_1 \in (0, \frac{2}{2+p})$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \tilde{c}_1 \log(8/\delta) m^{-\gamma_1},$$

where $\gamma_1 = \min\left\{2r\theta_1, \frac{1-\theta_1+2r\theta_1}{2}, \frac{2}{2+p} - (2-2r)\theta_1, \frac{2(1-p\theta_1)}{2+p}\right\}$.

(2) If $r \geq \frac{1}{2}$ in Assumption 1, taking $\lambda = m^{-\theta_2}$ with some $\theta_2 \in (0, \frac{2}{2+p})$,

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \tilde{c}_1 \log(8/\delta) m^{-\gamma_2},$$

where $\gamma_2 = \min\left\{\theta_2, \frac{1}{2}, \frac{2}{2+p} - \theta_2\right\}$.

4. Theoretical Analysis: Remark

- Theorem 1 provides the upper bound of generalization error to SSAM with Lipshitz continuous kernel.
- For $r \in (0, \frac{1}{2})$, as $s \to \infty$, we have $\gamma_1 \to \min\{2r\theta_1, \frac{1}{2} + (r - \frac{1}{2})\theta, 1 - 2\theta_1 + 2r\theta_1\}$.
- When $r \to \frac{1}{2}$ and $\theta_1 \to \frac{1}{2}$, the convergence rate $O(m^{-\frac{1}{2}})$ can be reached.
- For $r \geq \frac{1}{2}$, taking $\theta_2 = \frac{1}{2+p}$, we get the convergence rate $O(m^{-\frac{1}{2+p}})$.

4. Theoretical Analysis: Theorems

Theorem 2

Assume that $f_\rho^{(j)} \in \mathcal{H}^{(j)}$ for each $1 \leq j \leq d$. Take $\lambda = m^{-\frac{2}{2+3p}}$ in (1). For any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \leq \tilde{c}_2 \log(1/\delta) m^{-\frac{2}{2+3p}},$$

where $\tilde{c}_2$ is a positive constant independent of m, δ.

4. Theoretical Analysis: Theorems

Theorem 2
Assume that $f_\rho^{(j)} \in \mathcal{H}^{(j)}$ for each $1 \le j \le d$. Take $\lambda = m^{-\frac{2}{2+3p}}$ in (1). For any $0 < \delta < 1$, with confidence $1 - \delta$ we have

$$\mathcal{E}(\pi(f_{\mathbf{z}})) - \mathcal{E}(f_\rho) \le \tilde{c}_2 \log(1/\delta) m^{-\frac{2}{2+3p}},$$

where $\tilde{c}_2$ is a positive constant independent of m, δ.

- The result is about a special case when $f_\rho^{(j)} \in \mathcal{H}^{(j)}$.
- Under the strong condition on f_ρ, the convergence rate can be arbitrary close to $O(m^{-1})$ as $s \to \infty$.

5. Empirical Evaluation: Synthetic Data Setting

- Pairwise interaction setting: $k = 2, d = \binom{n}{2}$.

5. Empirical Evaluation: Synthetic Data Setting

- Pairwise interaction setting: $k = 2, d = \binom{n}{2}$.
- Each kernel on $\mathcal{X}^{(j)}$ is generated from Gaussian kernel.

5. Empirical Evaluation: Synthetic Data Setting

- Pairwise interaction setting: $k = 2, d = \binom{n}{2}$.
- Each kernel on $\mathcal{X}^{(j)}$ is generated from Gaussian kernel.
- Generate Data. We generate the n-dimensional input $x_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ with $x_{ij} = \frac{W_{ij} + \eta U_i}{1+\eta}$ and $n = 10$, where W and U are sampled from independent uniform distributions defined in $[-0.5, 0.5]$.

5. Empirical Evaluation: Synthetic Data Setting

- Pairwise interaction setting: $k = 2, d = \binom{n}{2}$.
- Each kernel on $\mathcal{X}^{(j)}$ is generated from Gaussian kernel.
- Generate Data. We generate the n-dimensional input $x_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$ and $n = 10$, where W and U are sampled from independent uniform distributions defined in $[-0.5, 0.5]$.
- Feature selection criterion. We make feature selection according to the magnitude of $\sum_{t=1}^{100} \hat{\alpha}_t^{(j)}$ for $j \in \{1, ..., 45\}$.

5. Empirical Evaluation: Synthetic Data Setting

- Pairwise interaction setting: $k = 2, d = \binom{n}{2}$.
- Each kernel on $\mathcal{X}^{(j)}$ is generated from Gaussian kernel.
- Generate Data. We generate the n-dimensional input $x_i = (x_{i1}, x_{i2}, ..., x_{in})^T$ with $x_{ij} = \frac{W_{ij}+\eta U_i}{1+\eta}$ and $n = 10$, where W and U are sampled from independent uniform distributions defined in $[-0.5, 0.5]$.
- Feature selection criterion. We make feature selection according to the magnitude of $\sum_{t=1}^{100} \hat{\alpha}_t^{(j)}$ for $j \in \{1, ..., 45\}$.
- Performance measure. The Precision@τ describes the number of truly informative features in the top-τ selected results.

5. Empirical Evaluation: Synthetic Data Result

Table 1: Precision@τ for feature selection

(a) Synthetic data I

f^*	(m,n,η)	τ	SSAM	COSSO
a	(100,10,0)	4	3.88	3.69
		5	3.92	3.81
		6	3.93	3.85
	(100,10,1)	4	3.37	2.58
		5	3.68	2.80
		6	3.82	2.91
b	(100,10,0)	1	0.97	1
		2	0.97	1
		3	0.97	1
	(100,10,1)	1	0.95	0.62
		2	0.95	0.65
		3	0.98	0.68
c	(100,10,0)	4	3.94	0.63
		5	3.97	0.68
		6	3.97	0.75
	(100,10,1)	4	3.69	0.84
		5	3.87	0.91
		6	3.92	0.94

(b) Synthetic data II

f^*	(m,n,η)	τ	SSAM	COSSO
e	(100,10,0)	2	1.05	0.73
		3	1.13	0.90
		4	1.20	0.90
	(100,10,1)	2	1.04	0.13
		3	1.10	0.16
		4	1.12	0.20
f	(100,10,0)	2	0.72	0.88
		3	0.93	1
		4	1.23	1
	(100,10,1)	2	1.90	0.94
		3	1.94	0.94
		4	1.95	0.97
g	(100,10,0)	3	2.94	2.98
		4	2.94	2.98
		5	2.94	3
	(100,10,1)	3	2.85	2.14
		4	2.85	2.40
		5	2.85	2.49

5. Empirical Evaluation: Real Data Results

Table: Average MSE on real-world benchmark data.

	SSAM	SALSA	COSSO	SpAM	Lasso
Insulin	**1.0146**	1.0206	1.1379	1.2035	1.1103
Skillcraft	**0.5432**	0.5470	0.5551	0.90545	0.6650
Airfoil	**0.4866**	0.5176	0.5178	0.9623	0.5199
Forestfire	**0.3477**	0.3530	0.3753	0.9694	0.5193
Housing	0.3787	**0.2642**	1.3097	0.8165	0.4452
CCPP	0.0694	0.0678	0.9684	**0.0647**	0.0740
Music	0.6295	**0.6251**	0.7982	0.7683	0.6349
Telemonit	0.0689	**0.0347**	5.7192	0.8643	0.0863

6. Discussion

- Computational complexity. It could be reduced by introducing distributed strategy as our future work.

6. Discussion

- Computational complexity. It could be reduced by introducing distributed strategy as our future work.
- To prove the feature selection consistency.

Thank You