Carnegie Mellon

Group Sparse Additive Models Junming Yin, Xi Chen, Eric. P. Xing School of Computer Science, Carnegie Mellon University

Motivation

- Group structures among covariates: SNPs within the same gene; genes that belong to the same pathway; and etc.
- Nonlinear covariate effects: nonlinear effects of genes on the phenotype.
- Group variable selection problem in the nonparametric setting.

Problem Setting

• The multivariate nonparametric regression problem:

$$Y = m(X_1, \dots, X_p) + \epsilon$$

- *n* data samples: $\{(\mathbf{x}^{(i)}, y^{(i)}) : \mathbf{x}^{(i)} \in \mathbb{R}^p, y^{(i)} \in \mathbb{R}, i = 1, ..., n\}.$
- A set of potentially overlapping groups of covariates is given a prior.
- Goal: estimate a sparse regression function

$$m(X_1,\ldots,X_p) = \mathbb{E}[Y \mid X_1,\ldots,X_p],$$

whose supports are a union of predefined groups.

Our Approach: GroupSpAM

- Additive Models (Hastie & Tibshirani, 1990): $m(X_1, \ldots, X_p) = \sum_{i=1}^{P} f_j(X_j)$
- Non-overlapping groups: $\bigcup_{g \in \mathcal{G}} g = \{1, \dots, p\}$ and $g \bigcap g' = \emptyset$
- Optimization (population version):

minimize
$$\frac{1}{2}\mathbb{E}\left[\left(Y - \sum_{j=1}^{p} f_j(X_j)\right)^2\right] + \lambda \sum_{g \in \mathcal{G}} \sqrt{|g|} \underbrace{\sqrt{\sum_{j \in g} \mathbb{E}\left[f_j^2(X_j)\right]^2}}_{\|\mathbf{f}_g\|}$$

subject to $\mathbb{E}[f_i(X_i)] = 0, j = 1, \dots, p$

• Challenges:

- Characterization of the thresholding condition for functional sparsity at the group level.
- Unlike group lasso (Yuan & Lin, 2006) and SpAM (Ravikumar et al., 2009), there is no closed-form solution to the stationary condition for each group of functions, in the form of a soft-thresholding operator.
- Extensions:
 - For overlapping groups (Jacob *et al.*, 2009), decompose each original component function into the sum of a set of latent functions and apply the functional group penalty to the decomposed functions.

Stationary and Thresholding Conditions

Theorem

Let $R_g = Y - \sum f_{j'}(X_{j'})$ be the partial residual after $g' \neq g j' \in g'$ removing all functions from group g. The stationary condition

of the problem with respect to $\mathbf{f}_q = \{f_i\}_{i \in q}$ while fixing all other groups $\{\mathbf{f}_{g'}: g' \neq g\}$ is

$$f_j + \sum_{\substack{j' \in g: j' \neq j}} \mathbb{E}[f_{j'} \mid X_j] - \mathbb{E}[R_g \mid X_j] + \lambda \sqrt{|g|} s_j = 0, \forall j \in g,$$

where $\mathbf{s}_g = \{s_j\}_{j \in g}$ is a vector of functions belonging to the subgradient of $\|\mathbf{f}_g\|$.

• We don't restrict the correlation structure of component functions in the same group:

$$P_j f_{j'} := \mathbb{E}[f_{j'}(X_{j'}) \mid X_j] \neq 0, \ \forall \ j \neq j$$

Theorem $f_i = 0 \ \forall j \in g \text{ if and only if}$

$$\sqrt{\sum_{j \in g} \mathbb{E}[(P_j R_g)^2]} \le \lambda \sqrt{|g|}.$$

• In the finite sample case, estimate the projection $P_j R_q$ by smoothing: $\widehat{\mathbf{P}}_{j} = \mathbf{S}_{j} \widehat{\mathbf{R}}_{q}, \forall j \in g,$

where $S_i \in \mathbb{R}^{n \times n}$ is a linear smoother matrix and $\widehat{\mathbf{R}}_q \in \mathbb{R}^n$ is the estimate of partial residuals after removing group g.

Backfitting Algorithm

Input: Data $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^{n}$, partition \mathcal{G} , and parameter λ . Initialize $\mathbf{f}_i = \mathbf{0} \forall j$; pre-compute smoother matrices $\mathbf{S}_i \forall j$. Cycle through group $g \in \mathcal{G}$ until convergence:

Compute the residual: $\widehat{\mathbf{R}}_g = \mathbf{y} - \sum_{g' \neq g} \sum_{j' \in g'} \widehat{\mathbf{f}}_{j'}$.

Estimate the group norm: $\widehat{\omega}_g = \sqrt{\frac{1}{n} \sum_{j \in q} \|\mathbf{S}_j \widehat{\mathbf{R}}_g\|^2}$. If $\widehat{\omega}_g \leq \lambda \sqrt{|g|}$,

Set $\hat{\mathbf{f}}_i = \mathbf{0}, \ \forall j \in g$.

Else,

Estimate $\hat{\mathbf{f}}_q$ by fixed point iteration,

 $\widehat{\mathbf{f}}_{g}^{(t+1)} = \left(\widehat{\mathbf{J}} + \frac{\lambda\sqrt{|g|}}{\|\widehat{\mathbf{f}}_{g}^{(t)}\|/\sqrt{n}}\mathbf{I}\right)^{-1}\widehat{\mathbf{Q}}\widehat{\mathbf{R}}_{g}.$

Output: Fitted functions $\hat{\mathbf{f}} = { \hat{\mathbf{f}}_j \in \mathbb{R}^n : j = 1, \dots, p }.$



Experiments

$$f' \in q$$

Simulation Study:

• Sample size n = 150 and dimension p = 200, 1000.

•
$$Y = \sum_{j=1}^{8} f_j(X_j) + \epsilon$$

- $X_j \sim \text{Uni}(-2.5, 2.5), \text{Corr}(X_j, X_k) = t^2/(1+t^2)$
- $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 2.02 \,(\text{SNR} = 3.0)$
- For group lasso and GroupSpAM, assume a group structure with blocks of 4 neighboring covariates.

Comparisons of difference methods in terms of support recovery and prediction accuracy

method	precision	recall	$#\hat{f}_1$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$#\hat{f}_8$	MSE	method	precision	n recall	$#\hat{f}_1$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$#\hat{f}_8$	MSE
GroupSpAM	1.00	1.00	100	100	100	100	100	100	100	100	7.22	GroupSpAM	0.89	0.99	100	100	100	100	98	98	98	98	7.26
SpAM	0.85	0.82	83	100	56	100	100	94	27	100	9.61	SpAM	0.71	0.46	88	75	0	83	100	0	4	15	8.48
COSSO	0.66	0.42	6	1	27	100	50	61	3	88	28.29	COSSO	0.23	0.41	11	61	22	90	76	10	10	47	13.72
GroupLasso	0.95	0.99	100	100	100	100	99	99	99	99	28.34	GroupLasso	0.13	0.12	14	14	14	14	11	11	11	11	26.19
		р	= 20)0, t	= 0									р	= 20	0, t	= 2						
method	precision	n recall	$#\hat{f_1}$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f_7}$	$#\hat{f}_8$	MSE	method	precisio	n recall	$#\hat{f_1}$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$#\hat{f_8}$	MSE
GroupSpAM	1.00	1.00	100	100	100	100	100	100	100	100	7.21	GroupSpAM	0.75	0.97	95	95	95	95	100	100	100	100	8.10
SpAM	0.86	0.68	49	91	25	100	100	71	7	97	11.66	SpAM	0.69	0.34	59	43	0	65	100	0	1	3	9.69
COSSO	0.01	0.97	93	100	97	100	100	100	84	100	36.59	COSSO	0.00	0.00	0	0	0	0	0	0	0	0	26.30
GroupLasso	0.93	0.97	98	98	98	98	97	97	97	97	29.49	GroupLasso	0.02	0.03	4	4	4	4	2	2	2	2	25.86

																			•				
method	precisior	n recall	$#\hat{f_1}$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$#\hat{f_8}$	MSE	method	precisior	n recall	$#\hat{f_1}$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$\#\hat{f}_8$, MSE
GroupSpAM	1.00	1.00	100	100	100	100	100	100	100	100	7.22	GroupSpAM	0.89	0.99	100	100	100	100	98	98	98	98	7.26
SpAM	0.85	0.82	83	100	56	100	100	94	27	100	9.61	SpAM	0.71	0.46	88	75	0	83	100	0	4	15	8.48
COSSO	0.66	0.42	6	1	27	100	50	61	3	88	28.29	COSSO	0.23	0.41	11	61	22	90	76	10	10	47	13.72
GroupLasso	0.95	0.99	100	100	100	100	99	99	99	99	28.34	GroupLasso	0.13	0.12	14	14	14	14	11	11	11	11	26.19
p = 200, t = 0										p :	= 20	0, t	= 2										
														-									
method	precision	ı recall	$\#\hat{f}_1$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$#\hat{f}_8$	MSE	method	precision	n recall	$#\hat{f}_1$	$#\hat{f}_2$	$#\hat{f}_3$	$#\hat{f}_4$	$#\hat{f}_5$	$#\hat{f}_6$	$#\hat{f}_7$	$\hat{\#f_8}$	MSE
method GroupSpAM	precision	n recall 1.00	$#\hat{f}_1$ 100	$\frac{\#\hat{f}_2}{100}$	$\frac{\#\hat{f}_3}{100}$	$\frac{\#\hat{f}_4}{100}$	$\frac{\#\hat{f}_5}{100}$	$\frac{\#\hat{f}_6}{100}$	$\frac{\#\hat{f}_7}{100}$	$\frac{\#\hat{f}_8}{100}$	MSE 7.21	method GroupSpAM	precision 0.75	n recall 0.97	$\frac{\#\hat{f}_1}{95}$	$\frac{\#\hat{f}_2}{95}$	$\frac{\#\hat{f}_3}{95}$	$\frac{\#\hat{f}_4}{95}$	$\frac{\#\hat{f}_5}{100}$	$\frac{\#\hat{f}_6}{100}$	$\frac{\#\hat{f}_7}{100}$	$\frac{\#\hat{f}_8}{100}$	MSE 8.10
method GroupSpAM SpAM	precision 1.00 0.86	n recall 1.00 0.68	$ #\hat{f_1} $ 100 49	$#\hat{f}_2$ 100 91	$\hat{\#f_3}$ 100 25	$#\hat{f}_4$ 100 100	$\#\hat{f}_5$ 100 100	$\frac{\#\hat{f}_{6}}{100}$ 71	$\frac{\#\hat{f}_{7}}{100}$	$#\hat{f}_8$ 100 97	MSE 7.21 11.66	method GroupSpAM SpAM	precision 0.75 0.69	n recall 0.97 0.34	$#\hat{f}_1$ 95 59	$#\hat{f}_2$ 95 43	$\#\hat{f}_{3}$ 95 0	$ \frac{\#\hat{f}_4}{95} $ 65	$#\hat{f}_5$ 100 100	$\frac{\#\hat{f}_6}{100}$	$\hat{\#f_7}$ 100 1	$\frac{\#\hat{f}_8}{100}$	MSE 8.10 9.69
method GroupSpAM SpAM COSSO	precision 1.00 0.86 0.01	n recall 1.00 0.68 0.97	$ #\hat{f_1} $ 100 49 93	$\#\hat{f}_{2}$ 100 91 100	$\#\hat{f}_{3}$ 100 25 97	$\#\hat{f}_4$ 100 100 100	$\#\hat{f}_5$ 100 100 100	$\#\hat{f}_{6}$ 100 71 100	$#\hat{f}_7$ 100 7 84	$\#\hat{f}_8$ 100 97 100	MSE 7.21 11.66 36.59	method GroupSpAM SpAM COSSO	precision 0.75 0.69 0.00	n recall 0.97 0.34 0.00	$\#\hat{f}_1$ 95 59 0	$\hat{\#f_2} = 95 \\ 43 \\ 0$	$\#\hat{f}_{3}$ 95 0 0	$\hat{\#f_4}$ 95 65 0	$\#\hat{f}_5$ 100 100 0	$\hat{\#f_6} \\ 100 \\ 0 \\ 0 \\ 0$	$\hat{\#f_7}$ 100 1 0	$rac{\#\hat{f}_8}{100} \ 3 \ 0$	MSE 8.10 9.69 26.30
method GroupSpAM SpAM COSSO GroupLasso	precision 1.00 0.86 0.01 0.93	n recall 1.00 0.68 0.97 0.97	$ # \hat{f}_1 $ 100 49 93 98	$\#\hat{f}_2$ 100 91 100 98	$\#\hat{f}_{3}$ 100 25 97 98	$\#\hat{f}_4$ 100 100 100 98	$\#\hat{f}_5$ 100 100 100 97	$\#\hat{f}_{6}$ 100 71 100 97	$\hat{\#f_7}$ 100 7 84 97	$\#\hat{f}_8$ 100 97 100 97	MSE 7.21 11.66 36.59 29.49	method GroupSpAM SpAM COSSO GroupLasso	precision 0.75 0.69 0.00 0.02	n recall 0.97 0.34 0.00 0.03	$\hat{\#f_1} = 95 = 59 = 0 = 4$	$\hat{\#f_2} = 95 \\ 43 \\ 0 \\ 4$	$\hat{\#f_3} = 95 \\ 0 \\ 0 \\ 4$	$\hat{\#f_4} = 95 = 65 = 0 = 4$	$\#\hat{f}_5$ 100 100 0 2	$rac{\#\hat{f}_6}{100} \ 0 \ 2$	$\hat{\#f_7}$ 100 1 0 2	$rac{\#\hat{f}_8}{100} \ 3 \ 0 \ 2$	MSE 8.10 9.69 26.30 25.86

p = 1000, t = 0

• True component functions (red) versus estimated component functions (blue)



Breast Cancer Data (van de Vijver et al., 2002):

- Sample size n = 295 tumors (metastatic v.s. non-metastatic and dimension p = 3,510 genes; reduce p to 300 top genes; reduce p to 30 by sure independence screening (Fan & Lv, 2008).
- Goal: to find a sparse set of genes that can discriminate two types of tumors.
- Genes in the same biological pathway are likely to perf the same functionality in the cell, hence more likely to b involved in the studied phenomenon in a group manner
- Each group consists of the set of genes in a pathway and groups are overlapping.

RAY AND STEPHANIE LANE **Center for Computational Biology**

Carnegie Mellon

Cor	mpo	nent Functions	Variance
$f_1(x)$	=	$-2\sin(2x)$	2.10
$f_2(x)$	—	x^2	3.47
$f_3(x)$	=	$\frac{2\sin(x)}{2{-}\sin(x)}$	0.98
$f_4(x)$	=	$\exp(-x)$	8.98
$f_5(x)$	=	$x^3 + 1.5(x - 1)^2$	14.57
$f_6(x)$	=	x	2.08
$f_7(x)$	=	$3\sin(\exp(-0.5x))$	0.80
$f_8(x)$	=	$-5\phi(x, 0.5, 0.8^2)$	3.76

p = 1000, t = 2



	fold	method	BER	#genes	#pathways
static)		GroupSpAM	0.353	55	196
enes	1	SpAM	0.362	91	266
		GroupLasso	0.384	44	238
e the		GroupSpAM	0.358	44	243
	2	SpAM	0.349	109	302
form		GroupLasso	0.365	56	248
		GroupSpAM	0.326	74	149
)C	3	SpAM	0.333	101	209
Ι.		GroupLasso	0.346	76	138
and					

BER: balanced error rate, the average of the errors in each tumor type.