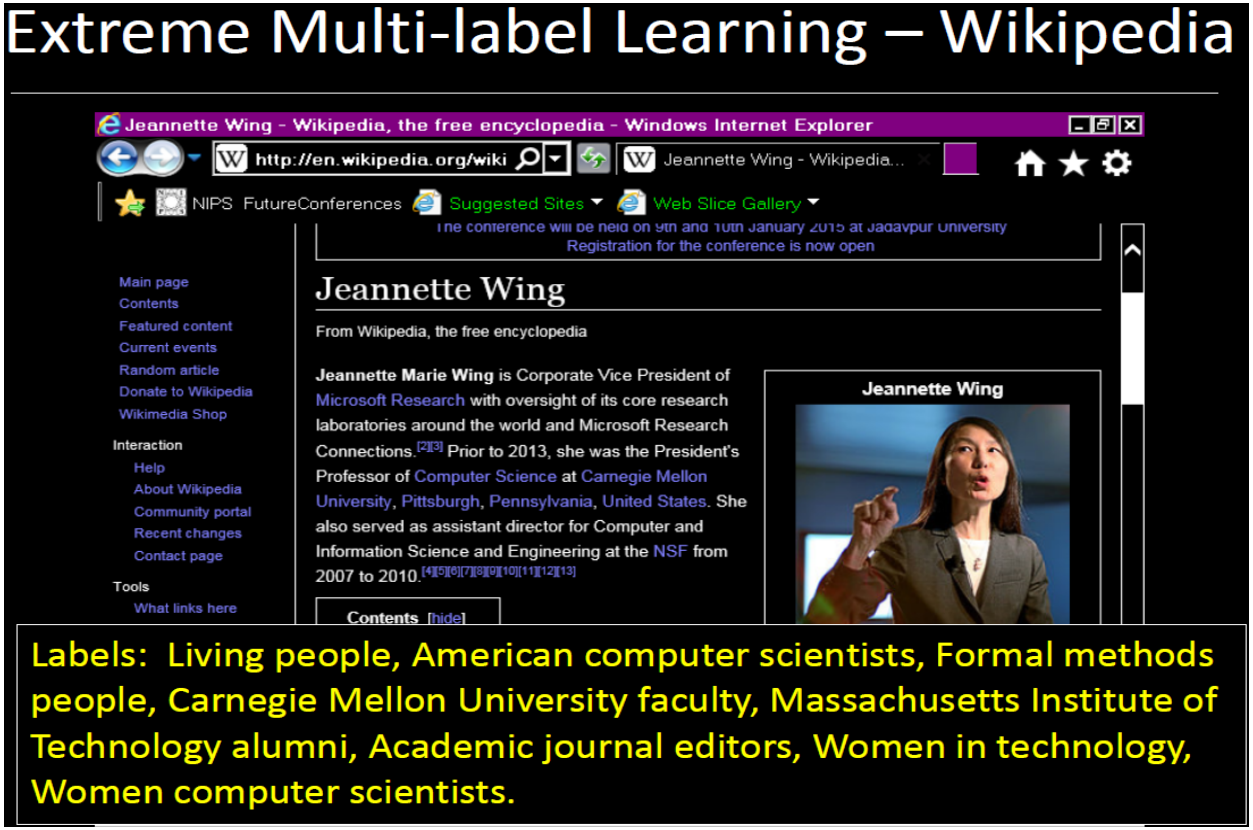
Sparse Extreme Multi-label Learning with Oracle Property

Introduction

- Extreme Multi-label classification

Each instance is associated with an extremely large number of labels.



Existing Methods

- Bhatia et al. (2015) :

SLEEC : learns a small ensemble of local distance preserving embeddings and shows great promise in extreme multi-label learning.

However, the statistical rate of convergence and oracle property of SLEEC remain less explored.

The Proposed Estimator

- Let V* represents the unknown sparse regression coefficient matrix and W denotes a noise matrix. We consider a multiple regression model as follows :

$$Z = V^* X + W$$

- We propose to estimate $\hat{V}$* by minimizing the following objective:

$$\hat{V} = \arg\min_V \| Z - VX \|_F^2 + \mu / 2 \| V \|_F^2 + \Re_\lambda (V)$$

where $\Re_\lambda (V)$ is a decomposable nonconvex regularization.

- Nonconvex penalty functions, such as smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and minimax concave penalty (MCP) (Zhang, 2010), have recently attracted much attention because they can eliminate the estimation bias and attain attractive statistical properties. This work takes SCAD and MCP penalties as the example.

Main Theory

- The oracle estimator is defined as: $\hat{V}_O = \underset{supp(V)\subseteq S^*}{\arg\min}\ \mathcal{L}(V)$

Theorem 1. *Suppose the nonconvex penalty* $\mathscr{P}_\lambda(V) = \sum_{(i,j)} p_\lambda(V_{(i,j)})$ *satisfies regularity conditions (i), (ii), (iii). We assume the oracle estimator* $\hat{V}_O$ *defined in Eq.(7) satisfies* $\min_{(i,j)\in S^*} |(\hat{V}_O)_{(i,j)}| \geq \nu$. *If* $\mu > \zeta$, $||X||_F \leq 1/n$, *and* V^* *satisfies* $||V^*||_\infty \leq 1/(\mu\sqrt{n})$, *we have*

(i) $\hat{V} = \hat{V}_O$.

(ii) $||\hat{V} - V^*||_F \leq \frac{4\sigma\sqrt{\varpi}+2\sqrt{s^*}}{\mu\sqrt{n}}$.

- Remark. Theorem 1 shows that our proposed estimator is identical to the oracle estimator under suitable conditions. This is a very strong result because we do not even have any oracle knowledge on the true support. Moreover, our proposed estimator is able to achieve the desirable statistical convergence rate for estimating V*.

Main Theory

Theorem 2. *We assume that* $|V^*_{(i,j)\in S_1^*}| \geq \nu$, *while* $|V^*_{(i,j)\in S_2^*}| < \nu$. *Suppose the nonconvex penalty* $\mathscr{P}_\lambda(V) = \sum_{(i,j)} p_\lambda(V_{(i,j)})$ *satisfies regularity conditions (i), (ii), (iii) and (iv). Given* $\mu > \zeta$, *for the estimator defined in Eq.(4) with regularization parameter* $\lambda = C\sqrt{\log m_1/nm_2}$ $(C > 0)$, *and* $\max_{(i,j)\in S^*\cup \bar{S}^*} |\nabla \mathcal{L}(V^*)_{(i,j)}| \leq \lambda$, *we have*

$$||\hat{V} - V^*||_F \leq \underbrace{\frac{C\sqrt{s_1^* \log m_1}}{(\mu-\zeta)\sqrt{nm_2}}}_{\Xi_1:|V^*_{(i,j)}|\geq\nu} + \underbrace{\frac{3C\sqrt{s_2^* \log m_1}}{(\mu-\zeta)\sqrt{nm_2}}}_{\Xi_2:|V^*_{(i,j)}|<\nu}$$

- Remark. The upper bound in Theorem 2 includes two parts corresponding to different magnitudes of the entries in V* : 1) the first part corresponds to the set of entries with larger magnitudes; and (2) the second part corresponds to the set of entries with smaller magnitudes. We are able to achieve the sharper convergence rate.

Precision and nDCG Results

Table 2. Precision@k (k=1,3,5) comparisons on three medium-sized data sets. The best results are in bold.

Datasets		CS	CPLST	ML-CSSP	1-vs-All	REML	FastXML	LEML	SLEEC	SML-SCAD	SML-MCP
Bibtex	P@1	58.87	62.38	44.98	62.62	65.13	63.42	62.54	65.08	66.43	**67.39**
	P@3	33.53	37.84	30.43	39.09	41.45	39.23	38.41	39.64	41.18	**42.86**
	P@5	23.72	27.62	23.53	28.79	30.12	28.86	28.21	28.87	30.25	**31.56**
Delicious	P@1	61.36	65.31	63.04	65.02	66.30	**69.61**	65.67	67.59	67.83	68.79
	P@3	56.46	59.95	56.26	58.88	61.73	64.12	60.55	61.36	63.45	**65.49**
	P@5	52.07	55.31	50.16	53.28	56.89	59.27	56.08	56.56	58.39	**60.56**
Mediamill	P@1	83.82	83.35	78.95	83.57	86.37	84.22	84.01	87.82	**89.56**	88.32
	P@3	67.32	66.18	60.93	65.50	73.97	67.33	67.20	73.45	**74.46**	73.89
	P@5	52.80	51.46	44.27	48.57	59.53	53.04	52.80	59.17	**60.53**	59.86

Table 3. nDCG@k (k=1,3,5) comparisons on three medium-sized data sets. The best results are in bold.

Datasets		CS	CPLST	ML-CSSP	1-vs-All	REML	FastXML	LEML	SLEEC	SML-SCAD	SML-MCP
Bibtex	nDCG@1	58.87	62.38	44.98	62.62	65.13	63.42	62.54	65.08	66.43	**67.39**
	nDCG@3	52.19	57.63	44.67	59.13	60.01	59.51	58.22	60.47	61.02	**61.23**
	nDCG@5	53.25	59.71	47.97	61.58	62.46	61.70	60.53	62.64	62.89	**63.04**
Delicious	nDCG@1	61.36	65.31	63.04	65.02	66.30	**69.61**	65.67	67.59	67.83	68.79
	nDCG@3	57.66	61.16	57.91	60.43	62.65	65.47	61.77	62.87	63.95	**66.76**
	nDCG@5	54.44	57.80	53.36	56.28	59.10	61.90	58.47	59.28	60.12	**62.13**
Mediamill	nDCG@1	83.82	83.35	78.95	83.57	86.73	84.22	84.01	87.82	**89.56**	88.32
	nDCG@3	75.29	74.21	68.97	73.84	82.67	75.41	75.23	81.50	**83.84**	82.35
	nDCG@5	71.92	70.55	62.88	68.18	78.32	72.37	71.96	79.22	**81.32**	80.63

Our proposed methods, SML-SCAD and SML-MCP, achieve the best results on all data sets.

Conclusion

- In this paper, we present a unified framework for SLEEC with nonconvex penalty.
- Our theoretical results show that our proposed estimator enjoys oracle property, and achieves an attractive statistical convergence rate.
- In addition, we can obtain a sharper convergence rate when a certain condition on the magnitude of the entries in the underlying model is imposed.
- Numerical experiments support our theoretical results and demonstrate the effectiveness of the proposed method.

Thank You!