Large-scale Multi-label Learning with Missing Labels

Hsiang-Fu Yu, Prateek Jain, Purushottam Kar, and Inderjit Dhillon
Department of Computer Science, University of Texas at Austin

Multi-label Learning

Setting:
- Data points: $(\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{y} \in \{0,1\}^L$.
- $f(\boldsymbol{x}; Z)$ parameterized by Z:
  $f(\boldsymbol{x}; Z) = Z^T\boldsymbol{x}$, $Z \in \mathbb{R}^{d\times L}$
- Applications: image/video annotation, query suggestions.

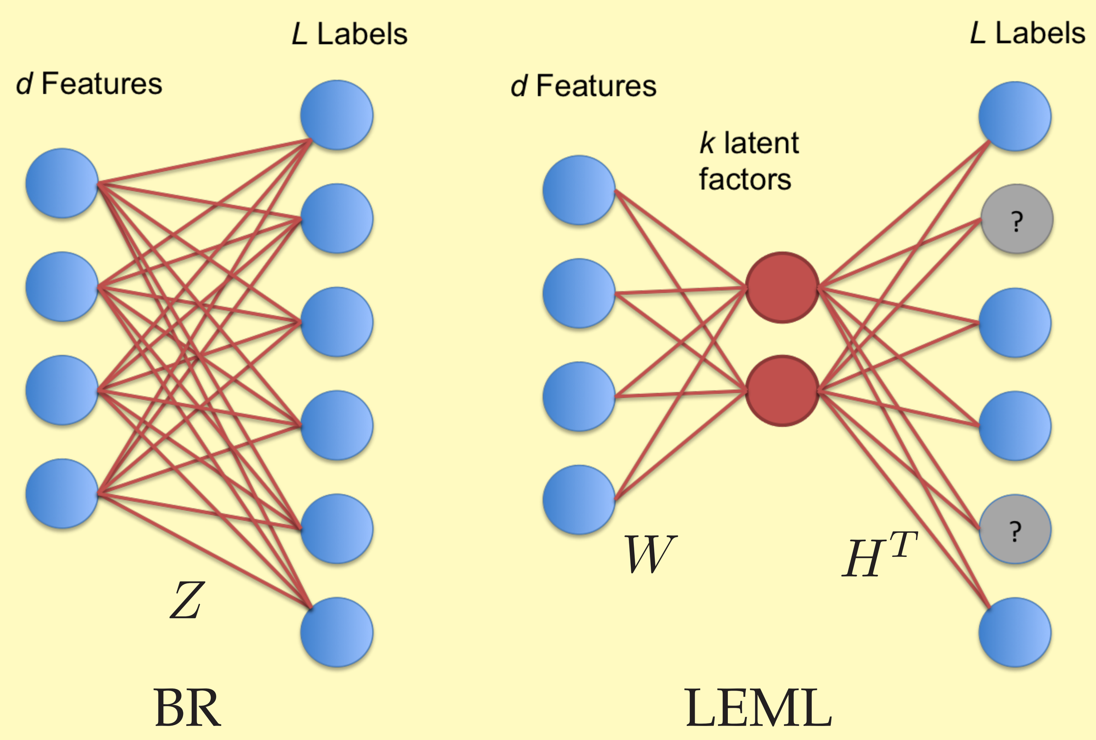
Existing Work
- Focus on Small L, label correlations
- Label-space reduction techniques
- Binary Relevance

Limits and challenges
- Scaling to extreme large L: storage and computation costs
- Working in presence of missing labels

Our Proposed Solution: LEML

What is LEML

Low-rank **E**RM for **M**ulti-Label **L**earning
- Learning model: $f(\boldsymbol{x}; Z) = Z^T\boldsymbol{x}$
- Low rank model: $Z = WH^T$
  Latent factors $\Leftarrow$ label correlations
  Computational + space savings
- ERM framework:
  Supports various loss functions
  Handles missing labels



$$\hat{Z} = \arg \min_{Z \in \mathbb{R}^{d\times L}} J_\Omega(Z) \equiv \lambda \cdot r(Z) + \hat{\mathcal{L}}_\Omega(Z),$$
$$s.t. \ \operatorname{rank}(Z) \le k.$$

- Low-rank constrained regularization:
$$\operatorname{rank}(Z) \le k \ \text{ and } \ r(Z) = \|Z\|_{tr}$$
- Empirical risk on $\Omega = \{(i,j) : Y_{ij} \ne ?\}$:
$$\hat{\mathcal{L}}_\Omega(Z) = \sum_{(i,j)\in\Omega} \ell(Y_{ij}, f^j(\boldsymbol{x}_i; Z))$$

LEML: Motivation

Why **L**ow-rank regularization?
- Significant **label correlations** when $L \gg 1$
  $\Rightarrow$ required # of parameters should $\ll d \times L$
- Rank constrained regularization:
  $\Rightarrow$ avoids overfitting
  $\Rightarrow$ computational benefits
- Why Trace norm: $\Rightarrow$ better at discovering latent structure.

Why **E**mpirical risk minimization (ERM)?
- Unified approach for various loss functions
- Extension for the presence of missing labels
- Generalization error analysis
- Many existing label embedding approaches are just a *special* case

LEML generalizes CPLST

A special case of LEML:
- Squared-L_2 loss: $\ell(\boldsymbol{y}, f(\boldsymbol{x}; Z)) = \|\boldsymbol{y} - Z^T\boldsymbol{x}\|_2^2$
- No regularization: $\lambda = 0$
- Fully observed: $\Omega = [n] \times [L]$

$$\hat{Z} = V_X \Sigma_X^{-1} M_k = \arg\min_{Z:\operatorname{rank}(Z)\le k} J(Z) \equiv \|Y - XZ\|_F^2,$$

where $X = U_X\Sigma_X V_X^T$, M_k is the best rank-k approximation of $M \equiv U_X^T Y$.

CPLST (Chen and Lin, NIPS 2012):
- $f_C(\boldsymbol{x}; Z_C) = Z_C^T\boldsymbol{x}$
- $Z_C = W_C H_C^T$, where W_C, H_C are minimizers of
$$\min \|XW - YH\|_F^2 + \|Y - YHH^T\|_F^2,$$
$$s.t. \quad H^T H = I_k$$
- $Z_C = W_C H_C^T = V_X \Sigma_X^{-1} M_k = \hat{Z}$

In fully observed case,

CPLST $\equiv$ LEML without regularization

Generalization Error Bounds

For general data distribution $\mathcal{D}$:
$$\mathcal{L}(\hat{Z}) \le \inf_{\|Z\|_{tr}\le\lambda} \mathcal{L}(Z) + \mathcal{O}\left(s\lambda\sqrt{\frac{1}{n}}\right) + \mathcal{O}\left(s\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$$

For any near isotropic data distribution $\mathcal{D}$:
$$\mathcal{L}(\hat{Z}) \le \inf_{\|Z\|_{tr}\le\lambda} \mathcal{L}(Z) + \mathcal{O}\left(s\lambda\sqrt{\frac{1}{nL}}\right) + \mathcal{O}\left(s\sqrt{\frac{\log\frac{1}{\delta}}{n}}\right)$$

- Frobenius norm cannot achieve this.

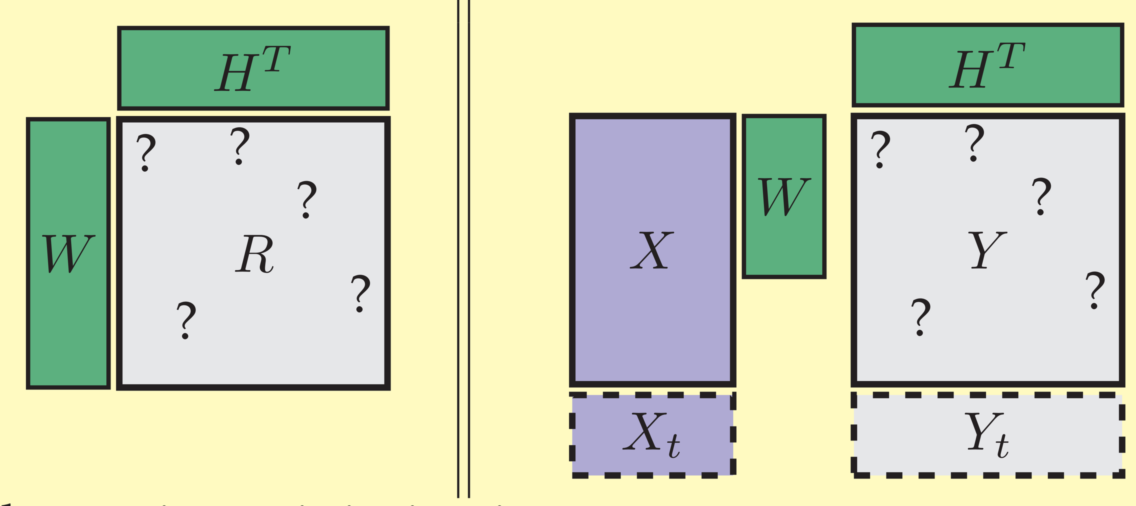
LEML: MF with Side Features

In LEML, $\operatorname{rank}(Z) \le k$ and $r(Z) = \|Z\|_{tr}$
- $Z = WH^T$, where $W \in \mathbb{R}^{d\times k}$ and $H \in \mathbb{R}^{L\times k}$
- $r(Z) = \|Z\|_{tr} = \min_{W,H} \frac{1}{2}\left(\|W\|_F^2 + \|H\|_F^2\right)$
- $f^j(\boldsymbol{x}_i; Z) = \boldsymbol{x}_i^T W\boldsymbol{h}_j$

Reformulation: $J_\Omega(Z) = J_\Omega(W, H) \equiv$
$$\sum_{(i,j)\in\Omega} \ell(Y_{ij}, \boldsymbol{x}_i^T W\boldsymbol{h}_j) + \frac{\lambda}{2}\left(\|W\|_F^2 + \|H\|_F^2\right)$$

Matrix Factorization
$R_{ij} \approx \boldsymbol{w}_i^T\boldsymbol{h}_j$

LEML: MF with row features
$Y_{ij} \approx \left(Z^T\boldsymbol{x}_i\right)_j = \boldsymbol{x}_i^T W\boldsymbol{h}_j$



Alternating minimization:
- For $t = 1, \ldots$
  - $H^{(t)} \leftarrow \arg\min_H J_\Omega(W^{(t-1)}, H)$ ...easy
  - $W^{(t)} \leftarrow \arg\min_W J_\Omega(W, H^{(t)})$ ...hard

Efficient Updates

$$W^{(t)} \leftarrow \arg\min_W J_\Omega(W, H^{(t)})$$

- Rows of W are not independent.
- Closed form solution: $\mathcal{O}(n\bar{L}\bar{d}dk^2 + d^3k^3)$ time, where $\bar{L} = |\Omega|/n$, $\bar{d} = nnz(X)/n$

Iterative solvers (e.g., conjugate gradient):
- *gradient* and *Hessian-vector multiplication*
- direct computation: $\mathcal{O}(n\bar{L}\bar{d}k)$ per operation

By exploiting the problem structure, we can
- reduce $\mathcal{O}(n\bar{L}\bar{d}k) \rightarrow \mathcal{O}(n\bar{L}k + n\bar{d}k)$ for a general smooth loss, and
- reduce $\mathcal{O}(nLk) \rightarrow \mathcal{O}(nnz(Y)k)$ for squared-L_2 with full labels
- See paper for the detailed algorithms
- Key techniques:

$$\underbrace{\sum_{(i,j)\in\Omega} C_{ij}\boldsymbol{a}_i\boldsymbol{b}_j^T}_{\mathcal{O}(|\Omega|dk)} = \underbrace{A^T C B}_{\mathcal{O}(|\Omega|k + ndk)}$$

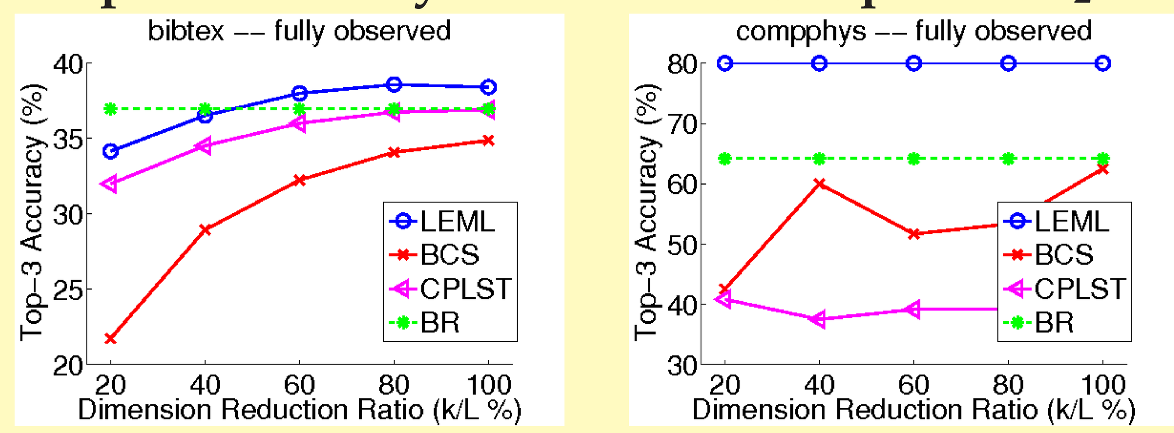
$$\underbrace{(B \otimes A)\,\boldsymbol{vec}(D)}_{\mathcal{O}(nLdk)} = \underbrace{\boldsymbol{vec}\left(ADB^T\right)}_{\mathcal{O}(ndk + nLk)}$$

Experimental Results

Datasets:

Dataset	Feature dimension d	$\bar{d}$	Label dimension L	$\bar{L}$	training n	test n
bibtex	1,836	68.74	159	2.40	4,880	2,515
autofood	9,382	143.92	162	15.80	155	38
compphys	33,284	792.78	208	9.80	161	40
eurlex	5,000	236.69	3,993	5.30	17,413	1,935
nus-wide	1,134	862.70	1,000	5.78	161,789	107,859
wiki	366,932	146.78	213,707	7.06	881,805	10,000

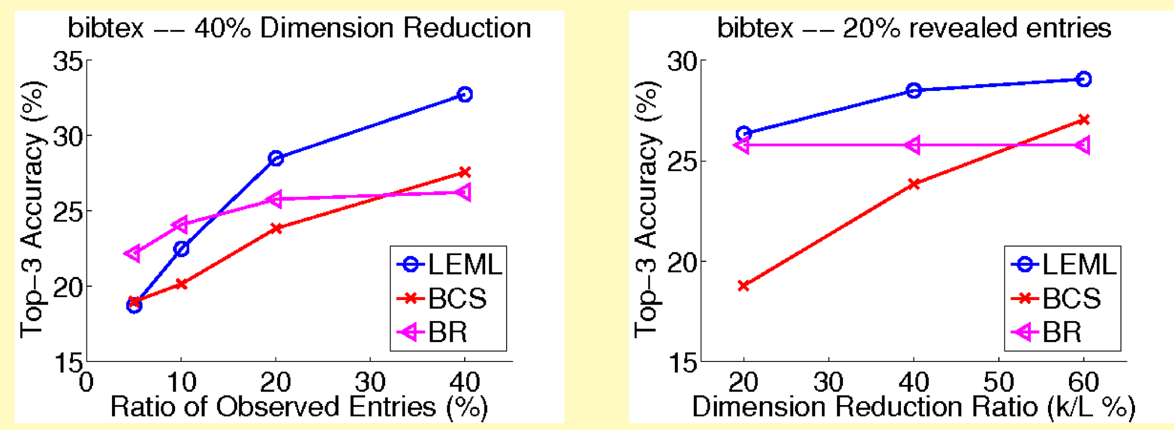
Comparison: Fully observation + Squared-L_2 loss



dataset	k	LEML time (s)	LEML top-1	LEML top-3	LEML AUC	WSABIE time (s)	WSABIE top-1	WSABIE top-3	WSABIE AUC
eurlex	250	**175**	**51.99**	**39.79**	**0.9425**	373	33.13	25.01	0.8648
	500	**487**	**56.90**	**44.20**	**0.9456**	777	31.58	24.00	0.8651
nus-wide	50	**574**	**20.71**	**15.96**	**0.7741**	4,705	14.58	11.37	0.7658
	100	**1,097**	**20.76**	**16.00**	**0.7718**	6,880	12.46	10.21	0.7597
wiki	250	**9,932**	**19.56**	14.43	**0.9086**	79,086	18.91	**14.65**	0.9020
	500	**18,072**	**22.83**	**17.30**	**0.9374**	139,290	19.20	15.66	0.9058

- CPLST suffers due to overfitting
- LEML can even beat BR for $k = L$ due to a trace norm regularizer.
- For wiki, MLR took 6 hours to get 0.9374, while WSABIE took 2 days to get 0.9058.

Comparison: Missing Labels + Squared-L_2 Loss



Top-3 Precision

dataset	Squared-L_2 LEML	Squared-L_2 BCS	Squared-L_2 BR	Logsitic LEML	Logsitic BR	Squared-Hinge LEML	Squared-Hinge BR
bibtex	**28.50**	23.84	25.78	25.79	**31.92**	18.97	**29.56**
autofood	**67.54**	35.09	62.28	**71.05**	53.51	**64.04**	61.40
compphys	**65.00**	35.83	31.67	**60.00**	28.33	**61.67**	31.67

Average AUC

dataset	Squared-L_2 LEML	Squared-L_2 BCS	Squared-L_2 BR	Logsitic LEML	Logsitic BR	Squared-Hinge LEML	Squared-Hinge BR
bibtex	**0.833**	0.781	0.808	0.839	**0.894**	0.781	**0.837**
autofood	**0.863**	0.632	0.817	**0.879**	0.771	**0.864**	0.817
compphys	**0.796**	0.644	0.745	**0.791**	0.729	**0.784**	0.745

References

1. Chen et al, Feature-aware label space dimension reduction for multi-label classification. NIPS 2012.
2. Weston et al, Large-scale image annotation: learning to rank with the joint word-image embeddings. Mach Learn, 81(1):21-35, 2010.