Latent Feature Lasso

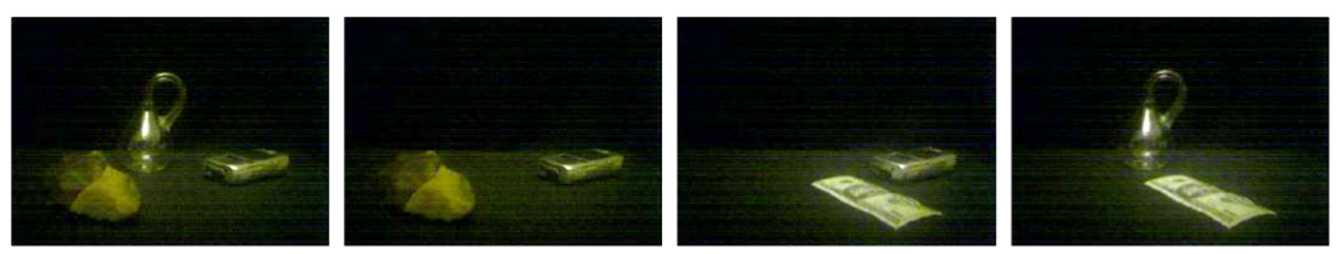
Ian E.H. Yen[1], Wei-Cheng Lee[2], Sung-En Chang[2], Arun S. Suggala[1], Shou-De Lin[2] and Pradeep Ravikumar[1]

[1]Carnegie Mellon University. [2]National Taiwan University

Abstract

- In this work, we propose a novel convex estimator (Latent Feature Lasso) for Latent Feature Model.
- To best of our knowledge, this is the first method with low-order polynomial runtime and sample complexity without restrictive assumptions on the data distribution for LFM.
- In experiments, the Latent Feature Lasso significantly outperforms other methods when there is a larger number of latent features.
- The method enjoys a runtime of $O(ND + DK^2)$ runtime per iter, more scalable than a typical $O(NDK^2)$ of existing approaches.

Latent Feature Models



11110 10110 10011 11001

- Latent Feature Model (LFM) is a generalization of Mixture Model, where each observation is an additive combination of latent features.

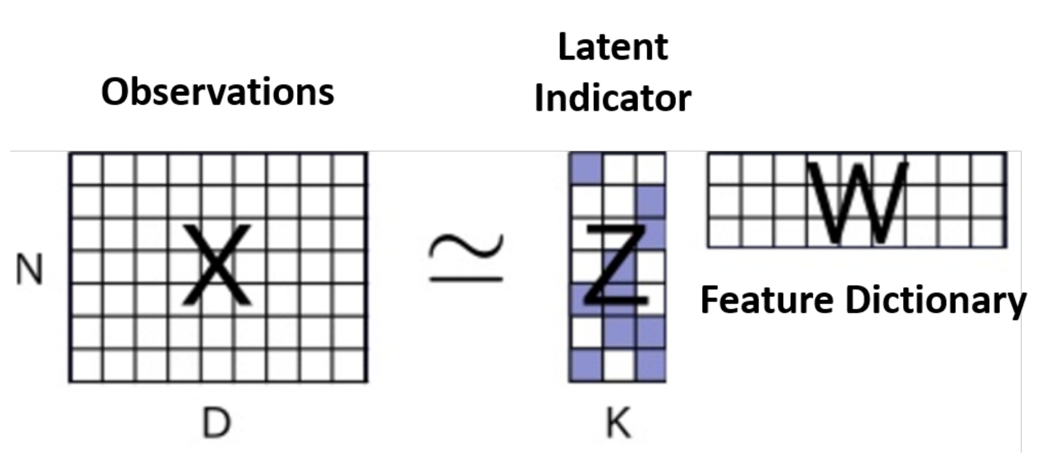
Discriminative	Multiclass Classification	Multilabel Classification
Generative	Mixture Model	Latent Feature Model

- In **Latent Feature Model**, each observation

$$\boldsymbol{x}_n = W^T \boldsymbol{z}_n + \epsilon_n$$

where $\boldsymbol{x}_n \in \mathbb{R}^D$: observation, $W \in \mathbb{R}^{K\times D}$: feature dictionary, $\boldsymbol{z}_n \in \{0,1\}^K$: binary latent indicators, and $\epsilon_n \in \mathbb{R}^D$: noise.

- **Mixture Model** is a special case with $\|\boldsymbol{z}_n\|_0 = 1$.



Related Works & Results

- **Goal:** Find dictionary $W_{K\times D}$ and latent indicators $Z: N \times K$ that best approximates observation $X: N \times D$.
- **Existing Approaches:**
  - MCMC, Variational (Indian Buffet Process): No finite-time guarantee.
  - Spectral Method (Tung 2014): $O(DK^6)$ sample complexity. ($z \sim$Ber(π), $x \sim N(W^T z, \sigma)$).
  - Matrix Factorization (Slawski et al., 2013): $O(NK2^K)$ runtime complexity for exact recovery (noiseless).
- **This Paper:**
  - A convex estimator — Latent Feature Lasso.
  - Low-order polynomial runtime and sample complexity.
  - No restrictive assumption on $p(X)$, even allows model mis-specification.

Convex Formulation via Atomic Norm

- Empirical Risk Minimization:

$$\min_{Z\in\{0,1\}^{N\times K}} \left\{ \min_{W\in\mathbb{R}^{K\times D}} \frac{1}{2N}\|X - ZW\|_F^2 + \frac{\tau}{2}\|W\|_F^2 \right\},$$

- Given Z, the dual problem w.r.t. W is:

$$\min_{M=ZZ^T\in\{0,1\}^{N\times N}} \underbrace{\left\{ \max_{A\in\mathbb{R}^{N\times D}} \frac{-1}{2N^2\tau} tr(AA^T M) - \frac{1}{N}\sum_{i=1}^{N} L^*(x_i, -A_{i,:}) \right\}}_{g(M)}.$$

- **Key insight:** the function is convex w.r.t. M.
- Enforce structure $M = ZZ^T$ via an atomic norm.
- Let $\mathcal{S} := \{k \mid \boldsymbol{z}_k \in \{0,1\}^N\}$. We define Atomic Norm:

$$\|M\|_{\mathcal{S}} := \min_{c\geq 0} \sum_{k\in\mathcal{S}} c_k \quad s.t. \quad M = \sum_{k\in\mathcal{S}} c_k \boldsymbol{z}_k \boldsymbol{z}_k^T.$$

- The Latent Feature Lasso estimator:

$$\min_M \; g(M) + \lambda\|M\|_{\mathcal{S}}.$$

- Equivalently, one can solve the estimator by

$$\min_{\boldsymbol{c}\in\mathbb{R}_+^{|\mathcal{S}|}} g\Big(\sum_{k\in\mathcal{S}} c_k \boldsymbol{z}_k \boldsymbol{z}_k^T\Big) + \lambda\|\boldsymbol{c}\|_1$$

Question: How to optimize with $|\mathcal{S}| = 2^N$ variables?

Greedy Coordinate Descent via MAX-CUT

- At each iteration, we find the coordinate of steepest descent:

$$j^* = \underset{j}{argmax} \; -\nabla_j f(c) = \underset{z\in\{0,1\}^N}{argmax} \; \langle -\nabla g(M), zz^T \rangle \tag{1}$$

which is a Boolean Quadratic problem similar to MAX-CUT:

$$\max_{\boldsymbol{z}\in\{0,1\}^N} \boldsymbol{z}^T C \boldsymbol{z}$$

- Can be solved to a $3/5$-approximation by roudning from a special type of SDP with $O(ND)$ iterative solver.

Active-Set Algorithm

0. $\mathcal{A} = \emptyset$, $c = 0$.

for $t = 1 \ldots T$ **do**

1. Find an approximate greedy atom zz^T by MAX-CUT-like problem:

$$\max_{z\in\{0,1\}^N} \langle -\nabla g(M), zz^T \rangle.$$

.

2. Add $\boldsymbol{z}\boldsymbol{z}^T$ to an active set $\mathcal{A}$.
3. Refine $\boldsymbol{c}_{\mathcal{A}}$ via Proximal Gradient Method on:

$$\min_{c\geq 0} g\Big(\sum_{k\in\mathcal{A}} c_k \boldsymbol{z}_k \boldsymbol{z}_k^T\Big) + \lambda\|\boldsymbol{c}\|_1$$

4. Eliminate $\{\boldsymbol{z}_k\boldsymbol{z}_k^T \mid c_k = 0\}$ from $\mathcal{A}$.

end for.

- Finding approximate greedy coordinate costs $O(ND)$ (via SDP).
- Evaluating $\nabla g(M)$: a least-square problem of cost $O(DK^2)$.
- Each iteration costs $\underbrace{O(ND)}_{\text{MAX-CUT}} + \underbrace{O(DK^2)}_{\text{Least-Square}}$

Runtime Complexity

MCMC	Variational	MF-Binary	BP-Means	Spectral	LatentLasso
$(NDK^2)T$	$(NDK^2)T$	$(NK)2^K$	$(NDK^3)T$	$ND + K^5 log(K)$	$(ND + K^2D)T$

Theoretical Results: Risk Bound

Let the population risk of a dictionary W be

$$r(W) := E[\min_{z\in\{0,1\}^K} \frac{1}{2}\|\boldsymbol{x} - W^T \boldsymbol{z}\|^2].$$

Let W^* be an optimal dictionary of size K, the algorithm outputs $\hat{W}$ with

$$r(\hat{W}) \leq r(W^*) + \epsilon$$

as long as

$$t = \Omega\Big(\frac{K}{\epsilon}\Big) \;\text{ and }\; N = \Omega\Big(\frac{DK}{\epsilon^3}\log\Big(\frac{RK}{\epsilon\rho}\Big)\Big).$$

- The result trades between risk and sparsity.
- No assumption on $\boldsymbol{x}$ except that of boundedness.
- The sample complexity is (quasi) linear to D and K.

Identifiability

Let $rank(\Theta^*) = K$. The decomposition $ZW = \Theta^*$ is unique if

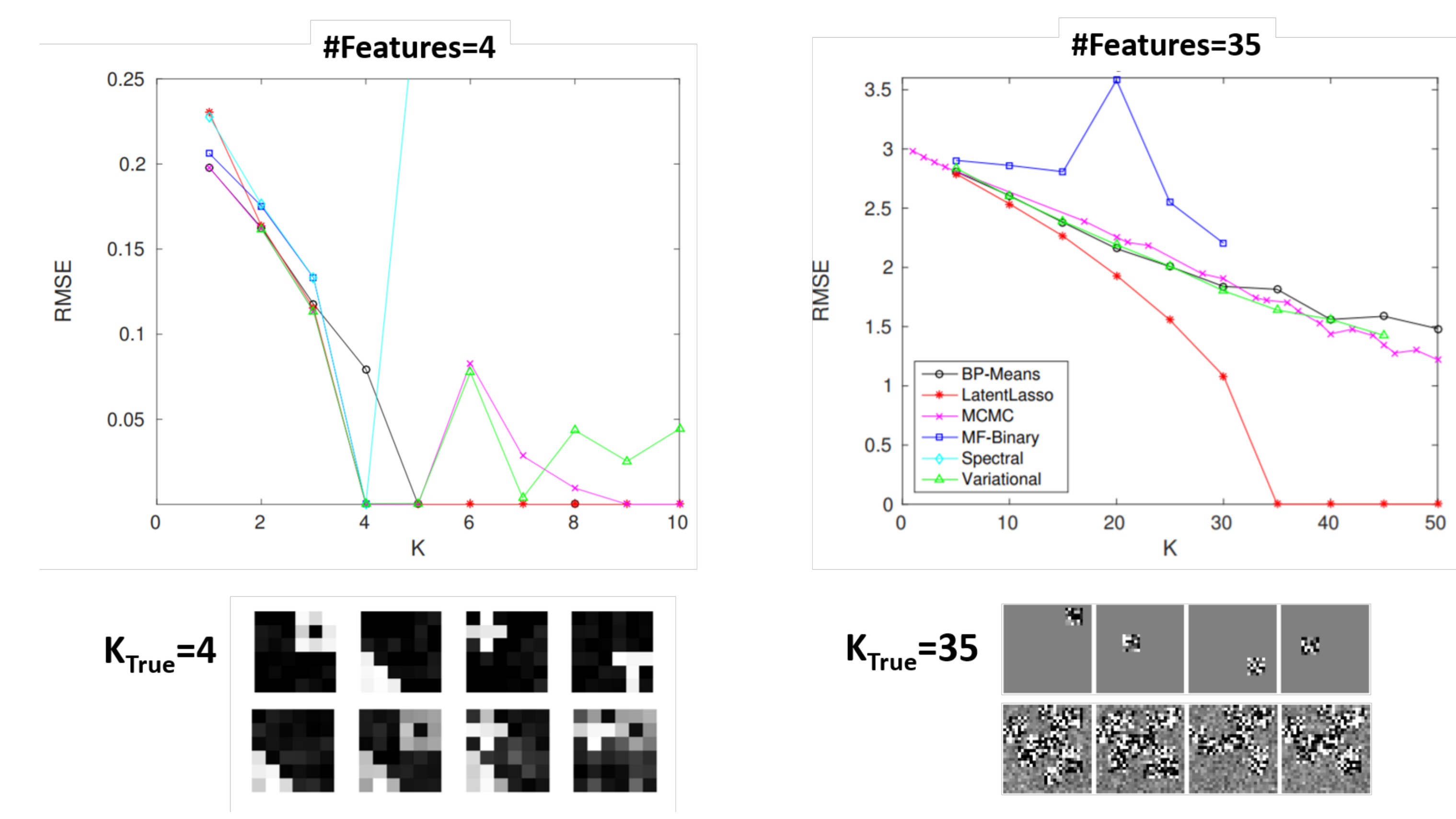
1. Z^*:$N \times K$ and W^*:$K \times D$ are both of rank K.
2. $span(Z^*) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K$.

Theoretical Results: Exact Recovery (noiseless)

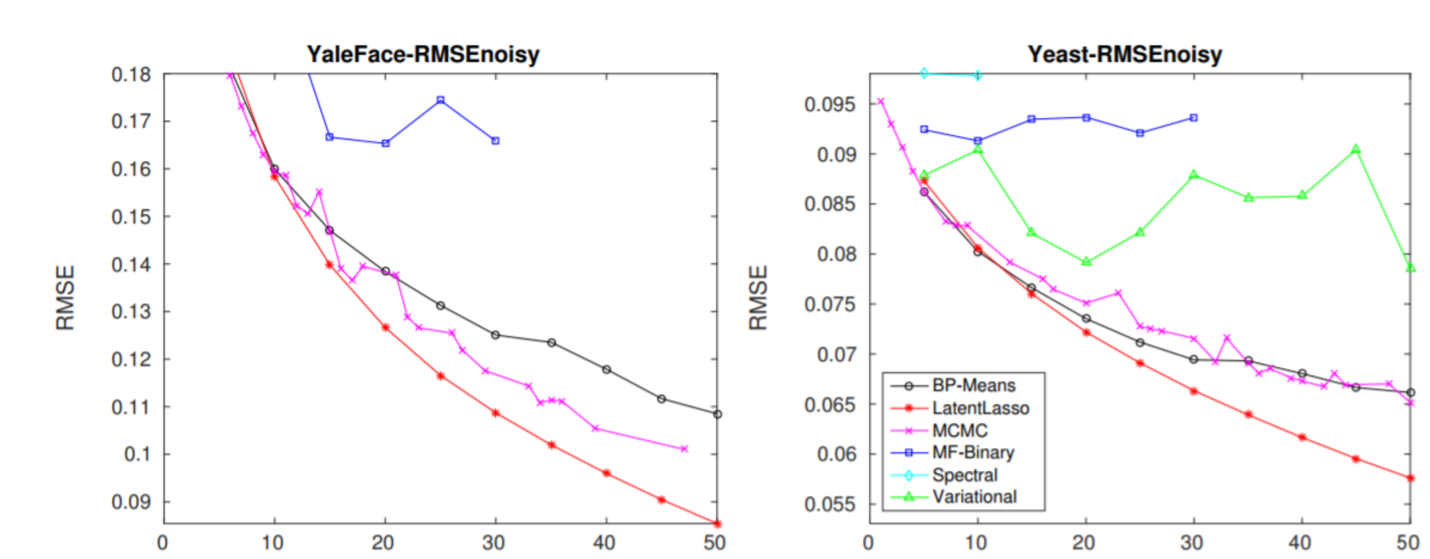
Let $X = Z^*W^*$, and $(Z_{\mathcal{A}}, W_{\mathcal{A}})$ be a solution of Latent Feature Lasso. If the identifiability holds and $W_{\mathcal{A}}$ has full row-rank:

$$\{Z_{:j}\}_{j\in\mathcal{A}} = \{Z^*_{:j}\}_{j=1}^K, \; \{W_{j,:}\}_{j\in\mathcal{A}} = \{W^*_{j,:}\}_{j=1}^K.$$

Experiments on Synthetic Data



Experiments on Real Data



Mail: {ianyen,pradeepr,inderjit}@cs.utexas.edu, zhongkai@ices.utexas.edu,chohsieh@ucdavis.edu