Latent Feature Lasso

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

\textsuperscript{1}Carnegie Mellon University \textsuperscript{2}National Taiwan University

Abstract

- In this work, we propose a novel convex estimator (Latent Feature Lasso) for Latent Feature Model.
- To the 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 $O(ND + D^2K)$ runtime per iter, more scalable than a typical $O(NDK^2)$ of existing approaches.

Convex Formulation via Atomic Norm

- Empirical Risk Minimization:
  $$\min_{z: \{1\}^N=0} \frac{1}{N} \sum_{i=1}^N L'(z_i, \hat{A}_i) + \frac{1}{2N} \|X - Wz\|^2_2 + \frac{1}{2} \|W\|^2_2.$$  
- Given $Z$, the dual problem w.r.t. $W$ is:
  $$\min_{W : \{1\}^K=0} -\frac{1}{2N} tr(\hat{A}A^T) - \frac{1}{N} \sum_{i=1}^N L'(z_i, \hat{A}_i).$$

- Key insight: the function is convex w.r.t. $M$.
- Enforce structure $M = Z^*Z$ via an atomic norm.
- Let $S := \{k \mid z_k \in \{0, 1\}^N\}$ We define Atomic Norm:
  $$\|M|_S := \sum_{k \in S} c_k s.t. M = \sum_{k \in S} c_k z_k^T z_k.$$  
- The Latent Feature Lasso estimator:
  $$\min_{g(M) + \lambda|\|M|_S|} \{ g(M) + \lambda|\|M|_S| \}.$$  
- Equivalently, one can solve the estimator by:
  $$\min_{\{k \mid z_k \in \{0, 1\}^N\}} \sum_{k \in S} c_k z_k^T z_k + \lambda|\|c||.$$  

Question: How to optimize with $|S| = 2^K$ variables?

Greedy Coordinate Descent via MAX-CUT

- At each iteration, we find the coordinate of steepest descent:
  $$f_j = \arg\max_{j \in \{1\}^N} -\nabla g(M)(z_j^*) = \arg\max_{j \in \{1\}^N} \max_{j \in \{1\}^N} -\nabla g(M)(z_j^*),$$

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

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

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

Active-Set Algorithm

0. $A = W, c = 0$.

for $t = 1,..., T$ do

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

$$\max_{z \in \{0, 1\}^N} -\nabla g(M)(z^*).$$

2. Add $zz^*$ to an active set $A$.

3. Refine $c_A$ via Proximal Gradient Method on:

$$\min_{c_A} \sum_{j \in A} g(z_j^* |z_j^*), \quad |\|c_A||.$$  

4. Eliminate $\{z_j^* | c_j = 0\}$ from $A$.

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 $O(ND + O(DK^2))$.
- MAX-CUT
- Least-Square

Runtime Complexity

- $O(ND)$
- $O(DK^2)$
- $O(K^2N)$
- $O(K^2)$
- $O(ND)$
- $O(K^2ND)$

Theoretical Results: Risk Bound

Let the population risk of a dictionary $W$ be:

$$r(W) := E\{ \min_{\{1\}^K=0} \frac{1}{2N} \|x - Wz\|^2_2 \}.$$  

Let $W^*$ be an optimal dictionary of size $K$, the algorithm outputs $W$ with:

$$r(W) \leq r(W^*) + \epsilon$$

as long as:

$$t = \Omega(K/(\epsilon^2)) \quad \text{and} \quad N = \Omega(DK^2 \log RK / \epsilon^2).$$

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

Identifiability

Let $\text{rank}(\{1\}^N) = K$. The decomposition $ZW = \delta^*$ is unique if $Z : N \times K$ and $W^T : K \times D$ are both of rank $K$.

Theoretical Results: Exact Recovery (noiseless)

Let $X = Z^*W^*$, and $\{Z_j, W_j\}$ be a solution of Latent Feature Lasso. If the identifiability holds and $W_j$ has full row-rank:

$$\{Z_j\}_{j \in A} = \{Z_j\}_{j \in A}^*, \quad \{W_j\}_{j \in A} = \{W_j\}_{j \in A}^*.$$  

Experiments on Synthetic Data

- #Features = 4
- #Features = 35

Experiments on Real Data

- $K_{max} = 4$
- $K_{max} = 35$