Latent Feature Lasso

Ian E.H. Yen*, Wei-Cheng Lee†, Sung-En Chang†, Arun Suggala*,
Shou-De Lin† and Pradeep Ravikumar*.

* Carnegie Mellon University
† National Taiwan University
Presenter: Wei-Cheng Lee

Outline

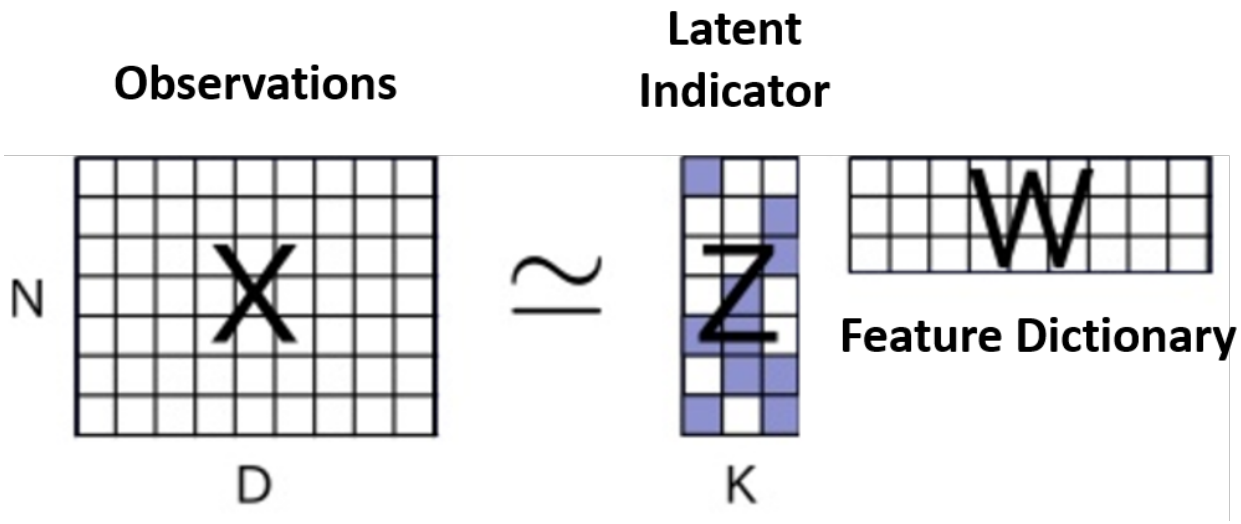
Latent Feature Models

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

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

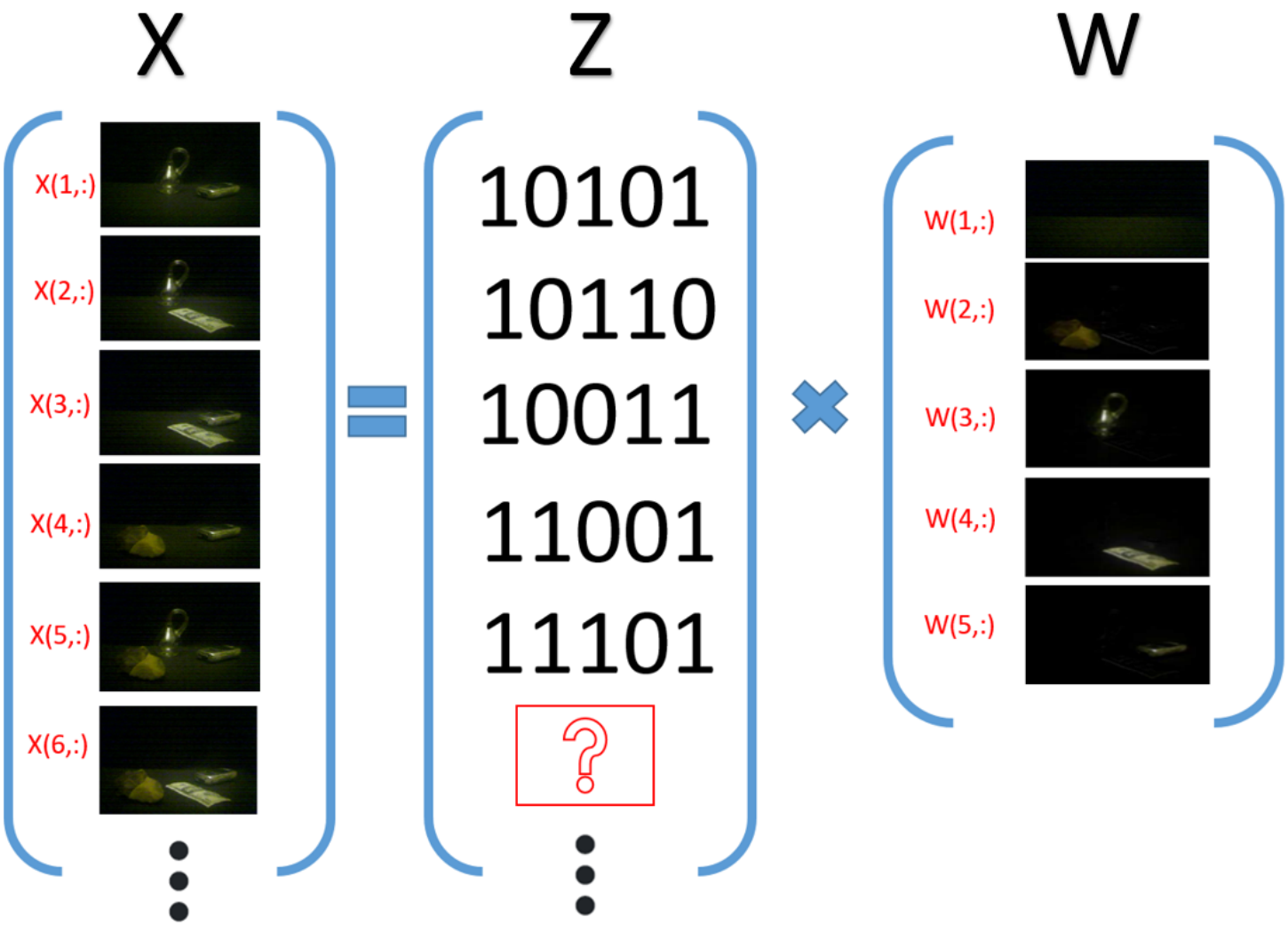
  where $\boldsymbol{x}_n \in \mathbb{R}^D$: observation, $W \in \mathbb{R}^{K \times D}$: feature dictionary, $\boldsymbol{z}_n \in \{0,1\}^K$: binary latent indicators, and $\boldsymbol{\epsilon}_n \in \mathbb{R}^D$: noise.

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



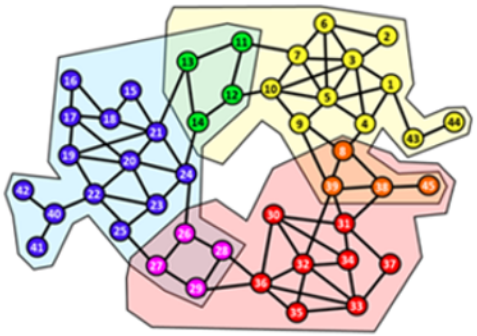
Latent Feature Model: Tabletop Dataset

- $X \in \mathbb{R}^{ND}$: dataset, $Z \in \{0,1\}^{NK}$: binary latent indicators, $W \in \mathbb{R}^{KD}$: feature dictionary, here K=5, D=picture size.

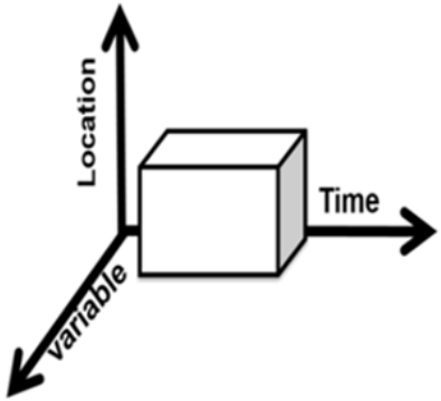


Latent Feature Model:Why Binary and Applications

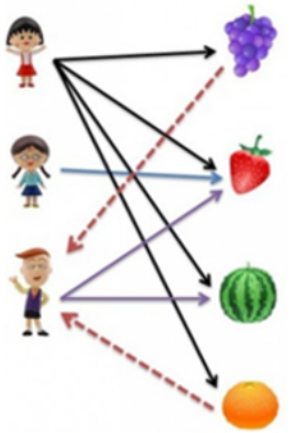
- Why binary? (interpretability, semi-supervision, and computational efficiency.)



Community Detection



Spatial-Temporal Factorization

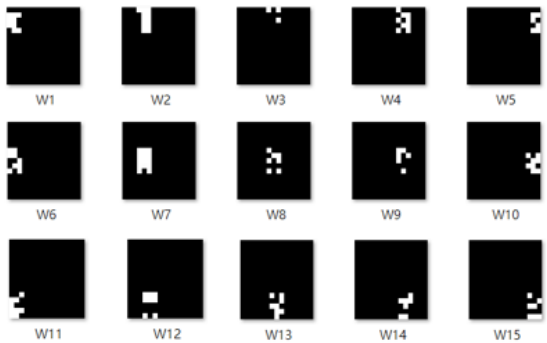


Recommendation

N	People	Location	Movies
D	People	Meteorological data	People
X value	0/1 for being friends or paper coauthorship	real value for temperature, precipitation, wind scale ...	Ratings
possible latent K	Groups, Affiliations	Weather events (typhoon, blizzard, cold current...)	Film genres

Let's Play a Game

- Under Latent Feature Model assumptions, Can you identify the latent features by your own eyes?



Latent Feature Models: Result Summary

- **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.

Outline

Previous Research

- Indian Buffet Process is a standard probability prior for the latent feature model.
- Previous works focus on using Gibs sampler or Variantional Inference to solve it.
- Spectral Method and BMF use different ways, but they have their own limitations.
- Sorry, I don't have time to survey previous works in details. So we just skip it QQ.

Outline

Latent Feature Model: Problem Formulation

- Goal: Empirical Risk Minimization:

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

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

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

- **Tricks:** Introducing dummy variable $E = ZW$, where A is its dual variable.
- **Key insight:** the function is convex w.r.t. $M = ZZ^T$.

Latent Feature Model: Estimator

- Let $\mathcal{S} := \{\boldsymbol{z}\boldsymbol{z}^T \mid \boldsymbol{z} \in \{0,1\}^N\}$.
- The "Latent-Feature" Atomic Norm:

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

- The Latent Feature Lasso estimator:

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

- Equivalently, one can solve the estimator by

$$\min_{\boldsymbol{c}\in\mathbb{R}_+^{|\mathcal{S}|}} \; g\left(\sum_{k=1}^{2^N} c_k \boldsymbol{z}_k \boldsymbol{z}_k^T\right) + \lambda\|\boldsymbol{c}\|_1$$

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

Outline

Greedy Coordinate Descent via MAX-CUT

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

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

  which is a Boolean Quadratic problem that can be reformulated to MAX-CUT:

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

- Can be solved to a 3/5-approximation (Nesterov) by rounding from a special type of SDP with $O(ND)$ iterative solver (Po-Wei Wangls 2016).

Greedy Coordinate Descent via MAX-CUT

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:

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

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

end for.

- Evaluating $\nabla g(M)$ requires solving a least-square problem of cost $O(DK^2)$.
- Each iteration costs $\underbrace{O(ND)}_{\text{MAX-CUT}} + \underbrace{O(DK^2)}_{\text{Least-Square}}$

Outline

Convergence Analysis

Given any reference solution c^*, the t-th iteration of the Greedy Algorithm satisfies

$$F(c^t) - F(c^*) \leq \frac{2\gamma\|c^*\|_1^2}{\mu^2}\left(\frac{1}{t}\right) + \underbrace{\frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1}_{\Delta(\lambda)},$$

$\mu = 3/5$ is the approximation ratio given by the MAX-CUT-like problem and γ is the Lipschitz-continuous parameter of $\nabla_j f(c)$.

- $\Delta(\lambda)$ decreases with N when λ is chosen to trade off between bias and variance.

Risk Analysis

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 bounds

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

under probability $1 - \rho$ as long as

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

and λ τ chosen appropriately corresponding to N.

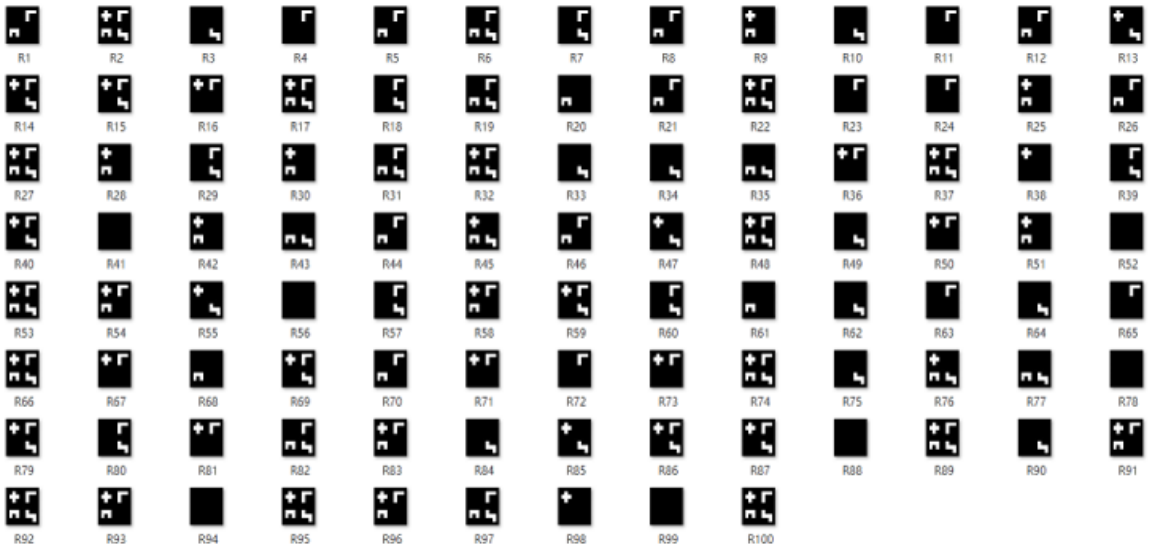
- The result trades between risk and sparsity.
- No assumption on $\boldsymbol{x}$ except that of boundedness.
- The sample complexity is (quasi) linear to D and K.

Outline

Results on Synthetic Data - k=4 case

- I will demonstrate some result on our algorithm under synthetic data. To point out that under same X, there may exist different solution pairs of (Z,W).
- In following example, we have n=100, k=4, d=50*50.

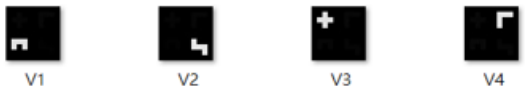
Each row of X



Each entry of Z follows Bernoulli(0.5)

True W

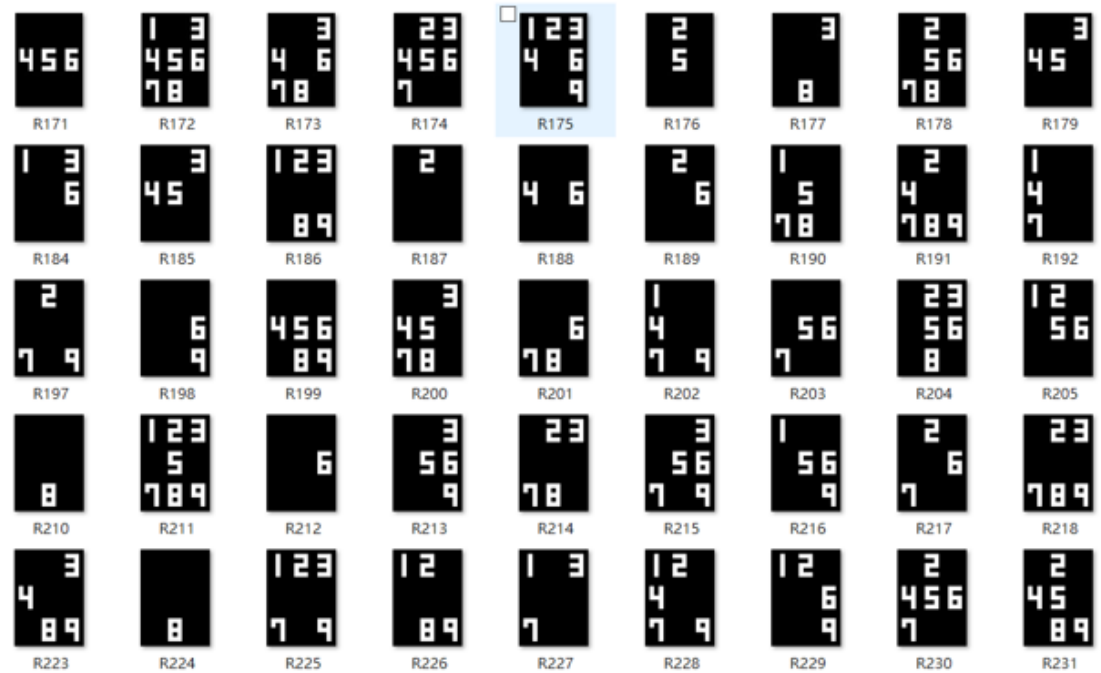
Our results



Results on Synthetic Data - k=9 Bernoulli(0.5)

- In this example, n=4500, k=9, d=105*75. Z follows Bernoulli(0.5) and W is number 1 to 9.

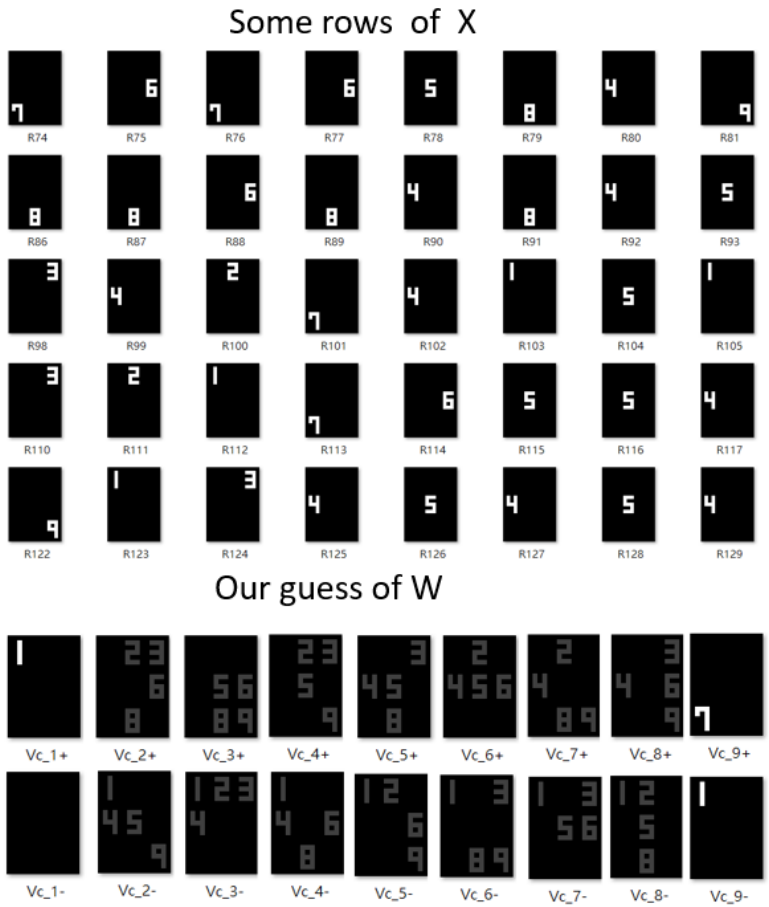
Some rows of X



Our guess of W

Results on Synthetic Data - k=9 Categorical(1/9)

- In this example, n=4500, k=9, d=105*75. Z follows Categorical(1/9) and W is number 1 to 9. We use gray-scale to represent value. $V = V^+ + V^-$. Observe that $R_{74} = V1 + V9$



Identifiability of Z and W

- Actually, even with same size of K, it is possible to have different Z and W. We have a theorem to restrict the solution of Z and W.

Theorem (Unique solution up to a Permutation)

Suppose $X \in \mathcal{R}^{N \times D}$, $Z^* \in \{0,1\}^{N \times K}$, $W^* \in \mathcal{R}^{K \times D}$, *given* $X = Z^* W^*$ *and rank(X)=K, If*

1. Z^* *is of rank* K.
2. $span(Z^*) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K$ ($\{Z^*_{:,j}\}_{j=1}^K$ *is a set containing all columns of* Z^*,$\{0,1\}^N$ *means all possible binary vector of length N)*

Consider another $Z \in \{0,1\}^{N \times K}$, $W \in \mathcal{R}^{K \times D}$ *such that* $X = ZW$ *and* Z *is also of rank K. Then there exists a permutation matrix* P *such that* $Z = Z^* P$ *and* $W = P^{-1} W^*$.

Is that two conditions achievable?

Proof.

Since $span(X) = span(Z^*) = span(Z)$. Therefore, from condition 2,

$$span(Z) \cap \{0,1\}^N \setminus \{0\} = span(Z) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K. \quad (1)$$

$\{Z^*_{:,j}\}_{j=1}^K = \{Z_{:,j}\}_{j=1}^K$ so there exist a permutation matrix P such that $Z = Z^*P$, then since the linear system $X = Z^*W_?$ has unique solution for $W_?$, we have $X = ZW = Z^*PW = Z^*W^*$. It follows that $W = P^{-1}W^*$. □

- It is a learning problem, we know X at first so we know its rank.
- In solving Z, we start from empty and add new columns to Z one by one by max-cut. It is easy to control the rank of Z to be K.
- The question is that what is the probability that condition 2 holds in general?

Previous results on random matrices

- I found the theorem in [Martin Slawski, Matthias Hein, Pavlo Lutsik Matrix factorization with Binary Components (Supplementary)]
- First theorem comes from [J. Kahn, J. Komlos, and E. Szemeredi. On the Probability that a ± 1 matrix is singular].
- Second theorem comes from [T. Tao and V. Vu. On the singularity problem of random Bernoulli matrices].

Theorem (1)

Let M be a random $m \times r$-matrix whose entries are drawn i.i.d from $\{-1,1\}$ each with probability $\frac{1}{2}$. There is a constant C so that if $r \leq m - C$ as $m \to \infty$

$$P(span(M) \cap \{-1,1\}^m) = \{\pm M_{:,1}, ..., \pm M_{:,r}\}) \geq 1 - (1 + o(1)))4\binom{r}{3}(\frac{3}{4})^m$$

Summary of Questions

Theorem (2)

Let M be a random m × r-matrix, r ≤ m, whose entries are drawn i.i.d from {-1,1} each with probability $\frac{1}{2}$.

$$P(span(M) \cap \{-1,1\}^m) = \{\pm M_{:,1}, ..., \pm M_{:,r}\}) \geq 1 - (\frac{3}{4} + o(1))^m \text{ as } m \to \infty$$

- Is it possible to use above two theorems to give a probability bound on our condition 2 given Z^* follows Bernoulli(0.5)?
- For Bernoulli(p)? or other distribution?
- Can we generate another constraint like condition 2 so that the solution is unique up to a permutation?
- In fact, it is quite strange to assume distribution of Z since we only know X. Is it possible to comes up with unique solution up to a permutation when we know the underlying distribution of X?

Outline

How to derive W and E and A

$$\min_{W \in \mathcal{R}^{K \times D}} \frac{1}{2N} \|X - Z_{\mathcal{A}} W\|_F^2 + \frac{\tau}{2} \|W\|_F^2$$

When $Z_{\mathcal{A}}$ is fixed, W has closed form solution

$$W^* = (Z_{\mathcal{A}}^T Z_{\mathcal{A}} + N\tau I)^{-1} Z_{\mathcal{A}}^T X$$

Given Lagrangian

$$L(W, E, A) = \frac{1}{2N} \|X - E\|_F^2 + \frac{\tau}{2} \|W\|_F^2 + < A, E - ZW >$$

$$0 \in \partial L \ (stationary)$$

$$E = ZW \ (primal \ \ feasibility)$$

Subgradient

A subgradient of a convex function $f : R^n \to R$ at a given point $x \in R^n$ is any $g \in R^n$ satisfies

$$\partial f(x) = \{g | f(y) \geq f(x) + g^T(y - x) \forall\ y\}$$

- $\partial f(x)$ always exists in this case.
- When f(x) is differentiable, $\partial f(x) = \{\nabla f(x)\}$
- x^* is optimal $\iff$ $0 \in \partial f(x^*)$

Proximal Gradient Descent-1

When convex f(x) is not totally differentiable but we can factorize it to

$$f(x) = g(x) + h(x)$$

where g(x) is convex/differentiable but h(x) is only convex. Same as what we do in gradient descent, we define

$$x^+ = x - t\nabla f(x)$$

but we need a proximal mapping operator defined as

$$prox_t(x) = \underset{z}{\text{argmin}} \frac{1}{2t}\|z - x\|^2 + h(z)$$

and in Proximal Gradient Descent, our update rule is given by

$$x^+ = prox_t(x - t\nabla g(x))$$

Proximal Gradient Descent-2

- We need to have an efficient way to compute the proximal operator itself (in closed form), otherwise we need to solve a quadratic optimization problem at every steps.
- Obtains convergence rate between subgradient methods and gradient methods.

Danskin's theorem

Suppose $\phi(x, z)$ is a continuous function

$$\phi : \mathcal{R}^n \times Z \to \mathcal{R}$$

where $Z \in \mathcal{R}^m$ is a compact set, and $\phi(x, z)$ is convex in x for every $z \in Z$ define

$$f(x) = \max_{z \in Z} \phi(x, z)$$

$$Z_0(x) = \{\bar{z} : \phi(x, \bar{z}) = \max_{z \in Z} \phi(x, z)\}$$

then f(x) is convex and

$$D_y f(x) = \max_{z \in Z_0(x)} \phi^{'}(x, z; y)$$

Max-Cut Solver-1

We can replace $X \succeq 0$ constraints with $X = V^T V$ for some $V \in \mathcal{R}^{k \times n}$ then $X_{ii} = 1$ translates to $\|v_i\| = 1$ leads to non-convex optimization problem

$$\min_{V \in \mathcal{R}^{k \times n}} < C, V^T V > \ \ subject\ to\ \|v_i\| = 1,\ \ i = 1, ..., n$$

Solve it by coordinate descent method, minimizing v_i that depends on $v_i^T (\sum_{j=1}^{n} C_{ij} v_j)$ since $\|v_i\| = 1$ we can assume that $C_{ii} = 0$ without affecting the solution, so the problem is equivalent to

$$\min_{v_i \in \mathcal{R}^k} v_i^T g \ \ subject\ to\ \|v_i\| = 1$$

The solution has closed form

$$v_i = \frac{-g}{\|g\|}$$

Max-Cut Solver-2

This is from "The Mixing method: coordinate descent for low-rank semi-definite programming (Po-Wei Wang) NIPS 2016"

Algorithm 1: The Mixing method

```
Initialize v_i randomly on a unit sphere;
while not yet converged do
    for i = 1, ..., n do
        v_i := normalize(−∑_{j=1}^{n} C_ij v_j);
    end
end
```

This way, we can initialize v_i on unit sphere and perform cyclic update over all the $i = 1, \ldots, n$ in closed-form. We called it the mixing method, because for each v_i it mixes and normalizes the remaining vectors v_j according to weight C_{ij}. Thus, in the case of sparse C (which is the normal case for any large data problem) the time complexity for updating all variable once is $O(k\#\text{nnz})$, which is significantly cheaper than the interior point method. However, the details for efficient computation differ depending on the precise nature of the SDP, so we will describe these in more detail in the subsequent application sections. A complete description of the generic algorithm is shown in Algorithm 1.