

Latent Feature Lasso

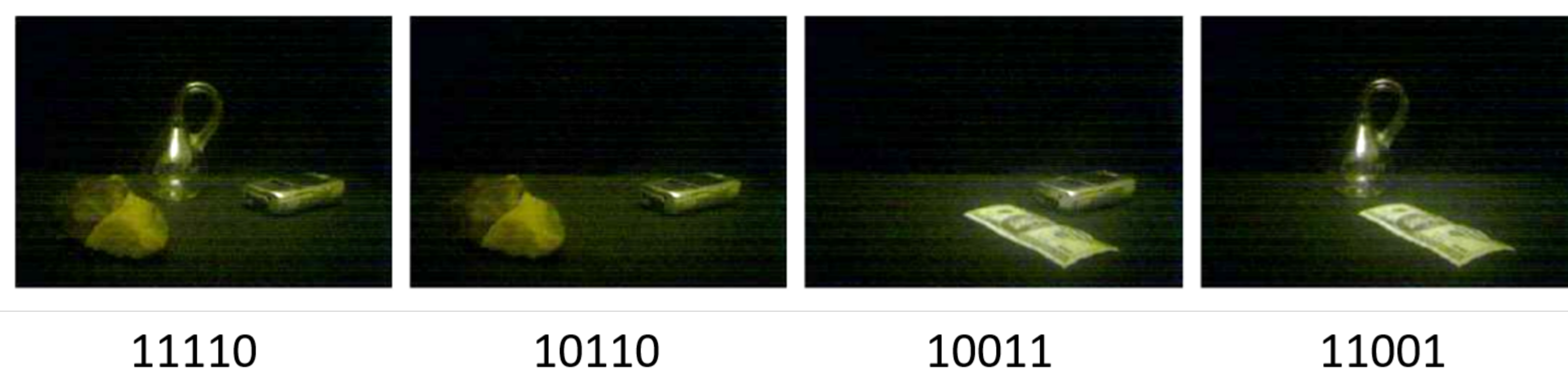
Ian E.H. Yen¹, Wei-Cheng Lee², Sung-En Chang², Arun S. Suggala¹, Shou-De Lin² and Pradeep Ravikumar¹

¹Carnegie Mellon University. ²National 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



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

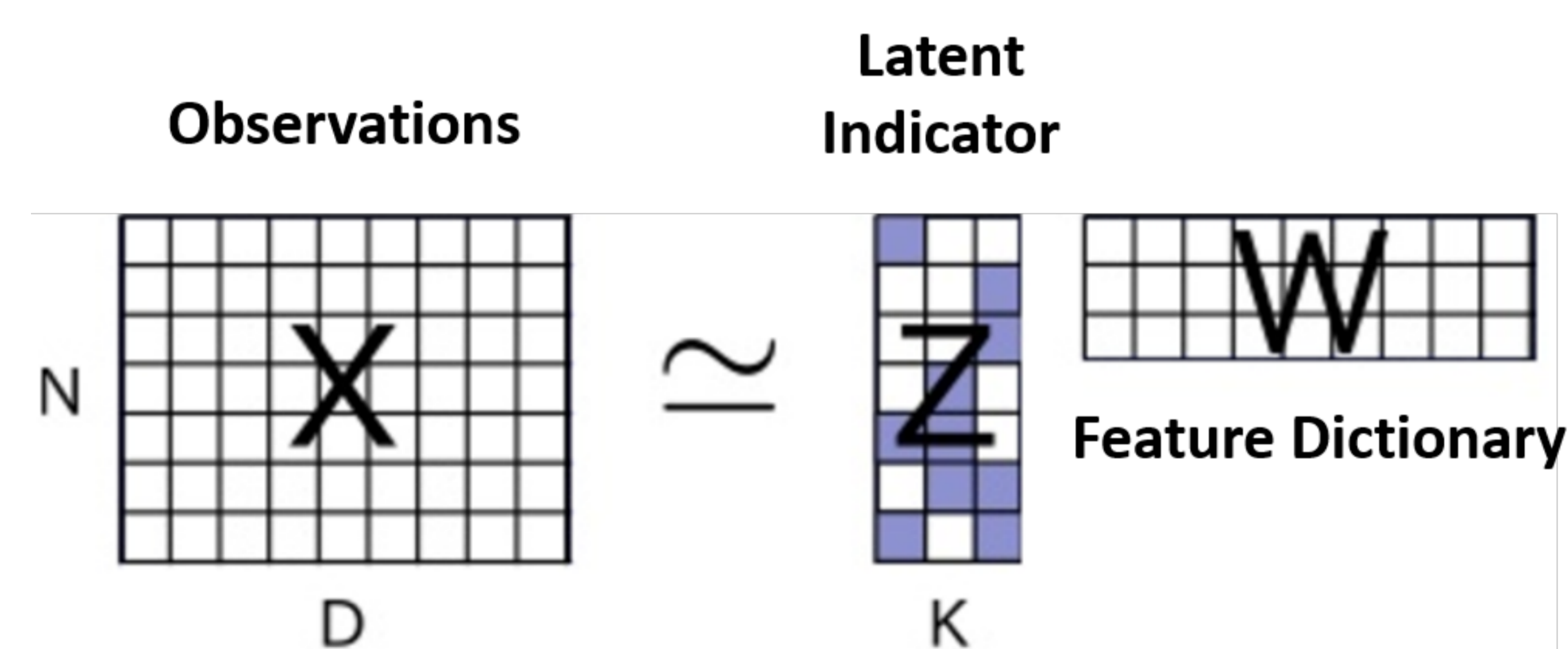
Discriminative	Multiclass Classification	Multilabel Classification
Generative	Mixture Model	Latent Feature Model

- In Latent Feature Model, each observation

$$\mathbf{x}_n = W^T \mathbf{z}_n + \mathbf{n}_n$$

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

- Mixture Model is a special case with $kz_n k_0 = 1$.



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. ($\mathbf{z} \sim \text{Ber}(\cdot)$, $\mathbf{x} \sim N(W^T \mathbf{z}; \cdot)$).
 - Matrix Factorization (Slawski et al., 2013): $O(NK^2K)$ 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:

$$\min_{Z \in \{0, 1\}^{N \times K}} \min_{W \in \mathbb{R}^{K \times D}} \frac{1}{2N} \sum_{i=1}^N \|X_i - Z_i W\|_2^2 + \frac{\lambda}{2} \|W\|_F^2$$

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

$$\min_{M \in \mathbb{R}^{N \times D}} \max_{A \in \mathbb{R}^{N \times D}} \frac{1}{2N^2} \text{tr}(AA^T M) - \frac{1}{N} \sum_{i=1}^N L(X_i; A_{i,:})$$

- Key insight:** the function is convex w.r.t. M .
- Enforce structure $M = ZZ^T$ via an atomic norm.

- Let $S := \{f_k \mathbf{z}_k^T\}_{k=1}^K$. We define Atomic Norm:

$$\|M\|_{k_2 S} := \min_{c \geq 0} \sum_{k=1}^K c_k \quad \text{s.t.} \quad M = \sum_{k=1}^K c_k \mathbf{z}_k \mathbf{z}_k^T$$

- The Latent Feature Lasso estimator:

$$\min_M g(M) + \|M\|_{k_2 S}$$

- Equivalently, one can solve the estimator by

$$\min_{c \geq 0} g\left(\sum_{k=1}^K c_k \mathbf{z}_k \mathbf{z}_k^T\right) + \sum_{k=1}^K c_k$$

Question: How to optimize with $\sum_{k=1}^K c_k = 2^N$ variables?

Greedy Coordinate Descent via MAX-CUT

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

$$j = \underset{j}{\text{argmax}} \quad r_j f(c) = \underset{j}{\text{argmax}} \quad h_j r g(M); \mathbf{z} \mathbf{z}^T \quad (1)$$

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

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

- 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. $A = \emptyset$, $c = 0$.

for $t = 1 \dots T$ do

- Find an approximate greedy atom $\mathbf{z} \mathbf{z}^T$ by MAX-CUT-like problem:

$$\max_{\mathbf{z} \in \{0, 1\}^N} h_j r g(M); \mathbf{z} \mathbf{z}^T$$

- Add $\mathbf{z} \mathbf{z}^T$ to an active set A .
- Refine c_A via Proximal Gradient Method on:

$$\min_{c \geq 0} g\left(\sum_{k \in A} c_k \mathbf{z}_k \mathbf{z}_k^T\right) + \sum_{k \in A} c_k$$

- Eliminate $f_{\mathbf{z}_k \mathbf{z}_k^T} c_k = 0$ from A .
- end for.

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

Runtime Complexity

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

Theoretical Results: Risk Bound

Let the population risk of a dictionary W be

$$r(W) := E\left[\min_{z \in \{0, 1\}^K} \frac{1}{2} \|X - z W\|_2^2\right]$$

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

$$r(\hat{W}) \leq r(W) +$$

as long as

$$t = \binom{K}{3} \quad \text{and} \quad N = \binom{DK}{3} \log\left(\frac{RK}{3}\right)$$

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

Identifiability

Let $\text{rank}(Z) = K$. The decomposition $ZW = X$ is unique if

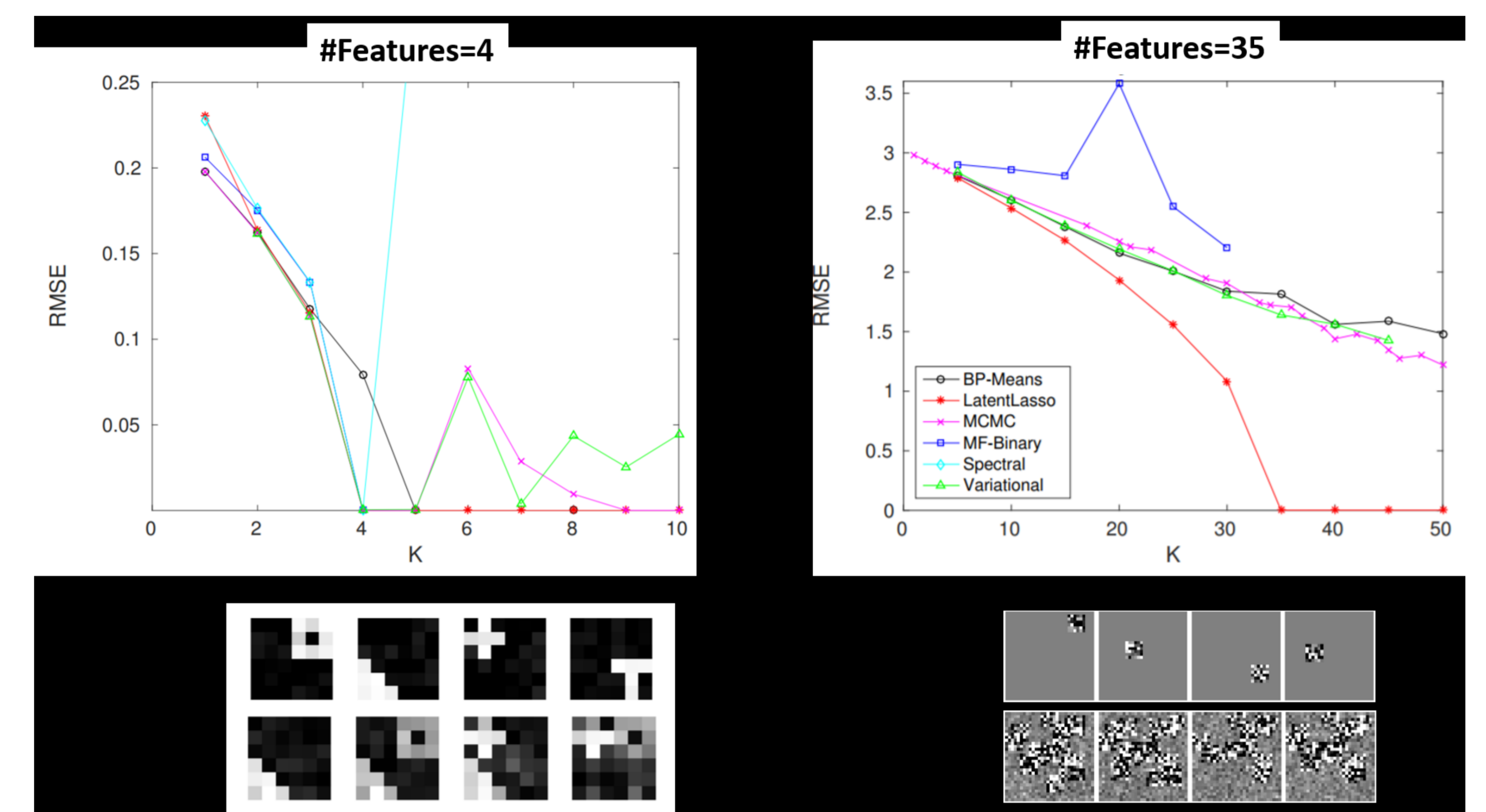
- $Z : N \times K$ and $W : K \times D$ are both of rank K .
- $\text{span}(Z) \setminus \{0\} \cap \text{span}(W) = \{0\}$.

Theoretical Results: Exact Recovery (noiseless)

Let $X = ZW$, and $(Z_A; W_A)$ be a solution of Latent Feature Lasso. If the identifiability holds and W_A has full row-rank:

$$f_{Z_A; j} g_{j2A} = f_{Z_A; j} g_{j-1}^K; \quad f_{W_A; j} g_{j2A} = f_{W_A; j} g_{j-1}^K$$

Experiments on Synthetic Data



Experiments on Real Data

