7. Appendix

7.1. Comparison on Time Complexity

The proposed LatentLasso algorithm runs significantly faster than other methods in our experiments. For example, on the Syn1 dataset (N=1000, D=1000, K=35), the runtime of LatentLasso is 398s, while MCMC, Variational, MF-Binary and BP-Means all take more than 10000s to obtain their best results reported in the Figures (and the implementation of Spectral Method we obtained from the authors has memory requirement that restricts K¡14). On the real data sets, we report only up to K=50 because most of the compared methods already took one day to train.

The complexity of each algorithm can be summarized in Table 2. The reason for the smaller runtime of LatentLasso is due to the decoupling of factor ND from the factor related to K, where the factor $O(ND)$ comes from the cost of solving a MAX-CUT-like problem using the method of (Boumal et al., 2016) or (Wang & Kolter, 2016), while the factor $O(K^2D)$ comes from the cost of solving a least-square problem given by (11) with the maintenance cost of Z^TZ amortized.

7.2. Proof for Theorem 1

Let $L(M)$ be a smooth function such that $\nabla L(M)$ is Lipschitz-continuous with parameter β, that is,

$$L(M') - L(M) - \langle \nabla L(M), M' - M \rangle \leq \frac{\beta}{2}\|M' - M\|_F^2.$$

Then

$$\nabla_j f(c) = z_j^T \nabla L(M) z_j$$

is Lipschitz-continuous with parameter γ, which is of order $O(1)$ when loss function $L(.)$ is an empirical average normalized by ND.

Let $\mathcal{A}$ be the active set before adding $\hat{j}$. Consider the descent amount produced by minimizing $F(c)$ w.r.t. the $c_{\hat{j}}$ given that $0 \in \partial_j F(c)$ for all $j \in \mathcal{A}$ due to the subproblem solved in the previous iteration. Let $j = \hat{j}$, for any η_j we have

$$\begin{aligned} F(c + \eta_j e_j) - F(c) &\leq \nabla_j f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &\leq \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \end{aligned}$$

Minimize w.r.t η_j gives

$$\begin{aligned} &\min_{\eta_j} \; F(c + \eta_j e_j) - F(c) \\ &\leq \min_{\eta_j} \; \mu \nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \\ &= \min_{\eta_k : k \notin \mathcal{A}} \; \sum_{k \notin \mathcal{A}} \left(\mu \nabla_k f(c)\eta_k + \lambda|\eta_k| \right) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &\leq \min_{\eta_k : k \notin \mathcal{A}} \; \mu \sum_{k \notin \mathcal{A}} \left(\nabla_k f(c)\eta_k + \lambda|\eta_k| \right) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &+ (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \end{aligned}$$

where the last equality is justified later in Lemma 1. For $k \in \mathcal{A}$, we have

$$0 = \min_{\eta_k : k \in \mathcal{A}} \; \mu \sum_{k \in \mathcal{A}} \left(\nabla_k f(c)\eta_k + \lambda|c_k + \eta_k| - \lambda|c_k| \right)$$

Combining cases for $k \notin \mathcal{A}$ and $k \in \mathcal{A}$, we can obtain a global estimate of descent amount compared to some optimal solution x^* as follows

$$\begin{aligned} &\min_{\eta_{\hat{j}}} \; F(c + \eta_{\hat{j}} e_{\hat{j}}) - F(c) \\ &\leq \min_{\eta} \; \mu \Big(\langle \nabla f(c), \eta \rangle + \lambda\|c + \eta\|_1 - \lambda\|c\|_1 \Big) \\ &+ \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 + (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\eta} \; \mu \Big(F(c + \eta) - F(c) \Big) + \frac{\gamma}{2} \left(\sum_{k \notin \mathcal{A}} |\eta_k| \right)^2 \\ &+ (1-\mu)\lambda \sum_{k \notin \mathcal{A}} |\eta_k| \\ &\leq \min_{\alpha \in [0,1]} \; \mu \Big(F(c + \alpha(c^* - c)) - F(c) \Big) + \frac{\alpha\gamma}{2}\|c^*\|_1^2 \\ &+ \alpha(1-\mu)\lambda\|c^*\|_1 \\ &\leq \min_{\alpha \in [0,1]} \; -\alpha\mu \Big(F(c) - F(c^*) \Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2 \\ &+ \alpha(1-\mu)\lambda\|c^*\|_1. \end{aligned}$$

It means we can always choose an α small enough to guarantee descent if

$$F(c) - F(c^*) > \frac{(1-\mu)}{\mu}\lambda\|c^*\|_1. \tag{23}$$

In addition, for

$$F(c) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1, \tag{24}$$

Table 2: Comparison of Time Complexity. (T denotes number of iterations)

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

we have

$$\min_{\eta_{\hat{j}}} F(c + \eta_{\hat{j}} e_{\hat{j}}) - F(c)$$
$$\leq \min_{\alpha \in [0,1]} -\frac{\alpha\mu}{2}\Big(F(c) - F(c^*)\Big) + \frac{\alpha^2\gamma}{2}\|c^*\|_1^2.$$

Minimizing w.r.t. to α gives the convergence guarantee

$$F(c^t) - F(c^*) \leq \frac{2\gamma\|c^*\|_1^2}{\mu^2}\frac{1}{t}.$$

for any iterate with $F(c^t) - F(c^*) \geq \frac{2(1-\mu)}{\mu}\lambda\|c^*\|_1$.

Lemma 1.

$$\min_{\eta_j} \mu\nabla_{j^*} f(c)\eta_j + \lambda|\eta_j| + \frac{\gamma}{2}\eta_j^2 \tag{25}$$
$$= \min_{\eta_k : k \notin \mathcal{A}} \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k + \lambda|\eta_k|\Big) + \frac{\gamma}{2}\left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \tag{26}$$

Proof. The minimization (34) is equivalent to

$$\min_{\eta_k : k\notin\mathcal{A}} \quad \sum_{k\notin\mathcal{A}}\Big(\mu\nabla_k f(c)\eta_k\Big)$$
$$s.t. \quad \left(\sum_{k\notin\mathcal{A}}|\eta_k|\right)^2 \leq C_1$$
$$\sum_{k\notin\mathcal{A}}|\eta_k| \leq C_2$$

and therefore is equivalent to

$$\min_{\eta_k : k\notin\mathcal{A}} \quad \mu\sum_{k\notin\mathcal{A}}\nabla_k f(c)\eta_k$$
$$s.t. \quad \sum_{k\notin\mathcal{A}}|\eta_k| \leq \min\{\sqrt{C_1}, C_2\}$$

which is a linear objective subject to a convex set and thus always has solution that lies on the corner point with only one non-zero coordinate η_{j^*}, which then gives the same minimum as (33). □

7.3. Proof of Theorem 2

Lemma 2. *Let $\mathcal{A}^* \in [\bar{K}]$ be a support set and $c^* := arg\min_{c:supp(c)=\mathcal{A}^*} F(c^*)$. Suppose $F(c)$ is strongly convex on $\mathcal{A}^*$ with parameter β. We have*

$$\|c^*\|_1 \leq \sqrt{\frac{2\|c^*\|_0(F(0) - F(c^*))}{\beta}}. \tag{27}$$

Proof. Since $supp(c^*) = \mathcal{A}^*$, and c^* is optimal when restricted on the support, we have $\langle , c^*\rangle = 0$ for some $\in \partial F(c^*)$. And since $F(c)$ is strongly convex on the support $\mathcal{A}^*$ with parameter β, we have

$$F(0) - F(c^*) = F(0) - F(c^*) - \langle , 0 - c^*\rangle$$
$$\geq \frac{\beta}{2}\|c^* - 0\|_2^2,$$

which gives us

$$\|c^*\|_2^2 \leq \frac{2(F(0) - F(c^*))}{\beta}.$$

Combining above with the fact for any c, $\|c\|_1^2 \leq \|c\|_0\|c\|_2^2$, we obtain the result. □

Since $F(0) - F(c^*) \leq \frac{1}{2N}\sum_{i=1}^N y_i^2 \leq 1$, from Theorem (1) and (27), we have

$$F(c^T) - F(c^*) \leq \frac{4\gamma\|c^*\|_0}{\beta\mu^2}\left(\frac{1}{T}\right) + \frac{2(1-\mu)\lambda}{\mu}\sqrt{\frac{2\|c^*\|_0}{\beta}}. \tag{28}$$

for any $c^* := arg\min_{c:supp(c)=\mathcal{A}^*} F(c)$.

7.4. Proof of Theorem 3

Before delving into the analysis of the *Latent Feature Lasso* method, we first investigate what one can achieve in terms of the risk defined in (1) if the *combinatorial version of objective* is solved. Let

$$f(x; W) := \min_{z\in\{0,1\}^K} \frac{1}{2}\|x - W^T z\|^2.$$

Suppose we can obtain solution $\hat{W}$ to the following empirical risk minimization problem:

$$\hat{W} := \underset{W\in\mathbb{R}^{K\times D}:\|W\|_F\leq R}{argmin} \quad \frac{1}{N}\sum_{i=1}^N f(x_i; W). \tag{29}$$

Then the following theorem holds.

Theorem 7. *Let W^* be the minimizer of risk (1) and $\hat{W}$ be the empirical risk minimizer (29). Then*

$$E[f(x;\hat{W})] - E[f(x;W^*)]$$
$$\leq \frac{3}{N} + \sqrt{\frac{DK\log(4R^2KN)}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}$$

with probability $1 - \rho$.

Proof Sketch. Let $E_N[f(x,W)]$ denote the empirical risk. We have

$$E[f(x;\hat{W})] - E[f(x;W^*)] \leq 2\left(\sup_{W\in\mathbb{R}^{K\times D}:\|W\|_F\leq R} |E[f(x;W)] - E_N[f(x;W)]|\right) \tag{30}$$

from error decomposition and $E_N[f(x,\hat{W})] \leq E_N[f(x,W^*)]$. Then by introducing a δ-net $\mathcal{N}(\delta)$ with covering number $|\mathcal{N}(\delta)| = \left(\frac{4R}{\delta}\right)^{DK}$, we have $\|\tilde{W} - W\|_F \leq \delta$ for some $\tilde{W} \in \mathcal{N}(\delta)$ and

$$P\left(\sup_{\tilde{W}\in\mathcal{N}(\delta)}\left|E[f(x;\tilde{W})] - E_N[f(x;\tilde{W})]\right| \leq \epsilon\right) \geq 1 - \left(\frac{4R}{\delta}\right)^{DK}\exp(-2N\epsilon^2). \tag{31}$$

Then since

$$\begin{aligned}2(f(x,\tilde{W}) - f(x,W)) &\leq \|x - \tilde{W}^T z^*\|^2 - \|x - W^T z^*\|^2 \\ &= z^{*T}(W - \tilde{W})x + \langle \tilde{W}\tilde{W}^T - WW^T, z^* z^{*T}\rangle \\ &\leq \|z^*\|_2\|W - \tilde{W}\|_F + 2R\|\tilde{W} - W\|_F\|z^*\|_2^2 \leq 3RK\|\tilde{W} - W\|_F\end{aligned}$$

we have

$$\begin{aligned}&\sup_{W:\|W\|_F\leq R}\left|E[f(x;W)] - E_N[f(x;W)]\right| \\ &\leq (3RK\delta) + \sup_{\tilde{W}\in\mathcal{N}(\delta)}\left|E[f(x;\tilde{W})] - E_N[f(x;\tilde{W})]\right| \\ &\leq 3RK\delta + \sqrt{\frac{DK}{2N}\log(\frac{4R}{\delta}) + \frac{1}{2N}\log(\frac{1}{\rho})}\end{aligned} \tag{32}$$

with probability $1-\rho$. Choosing $\delta = 1/(RKN)$ yields the result. □

Now we establish the proof of Theorem (3) for bounding risk of the *Latent Feature Lasso* estimator.

Proof. Let $Z^* \in arg\min_{Z\in\{0,1\}^{NK}} \frac{1}{N}\|X - ZW^*\|_F^2$ and $\mathcal{S}^*$ be the set of column index of $\mathbf{Z}$ with the same 0-1 patterns to columns in Z^*. Let c^* be indicator vector with $c^*_k = 1, k \in \mathcal{S}^*$ and $c^*_k = 0, k \notin \mathcal{S}^*$. We have

$$F(\bar{c}) \leq F(c^*) \leq E_N[f(x;W^*)] + \frac{\tau}{2}\|W^*\|_F^2 + \lambda\|c^*\|_1 \tag{33}$$

where $\bar{c} \in \underset{c:supp(c)=\mathcal{S}^*}{argmin} F(c)$. Then let (c,W) with $supp(c) = \hat{\mathcal{S}}$ be the output obtained from running T iterations of the greedy algorithm, we have

$$\begin{aligned}&E_N[f(x,D_cW)] + \frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \\ &= \frac{1}{2N}\sum_{i=1}^N \min_{z\in\{0,1\}^{\|c\|_0}}\|x_i - W^T D_c^T z\|^2 + \frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \\ &\leq F(c)\end{aligned} \tag{34}$$

Combining (33), (34) and (28), we obtain a bound on the *bias* and *optimization error* of the Latent Feature Lasso estimator

$$E_N[f(x,D_cW)] \leq F(c) \leq E_N[f(x;W^*)] + \underbrace{\frac{\tau}{2}\|W^*\|^2 + \lambda K}_{\text{regularize bias}} + \underbrace{\frac{2\gamma K}{\beta}\left(\frac{1}{T}\right) + \sqrt{\frac{2(1-\mu)K}{\mu\beta}}\lambda}_{\text{optimization error}} \tag{35}$$

To bound the estimation error, notice that the matrix $\hat{W} := D_cW$ is $\hat{K}\times D$ with $\hat{K} \leq T$. Furthermore, the descent condition $F(c) \leq F(0)$ guarantees that

$$\frac{\tau}{2}\|W\|_F^2 + \lambda\|c\|_1 \leq \frac{1}{N}\|X - 0\|^2 \leq 1$$

and thus $\|W\|_F^2 \leq 1/\tau$, $\|c\|_1 \leq 1/\lambda$.

Let $\mathcal{W}(T,\lambda,\tau) := \{\hat{W} \in (\mathbb{R}^{T\times D}) \mid \|\hat{W}\|_F \leq \sqrt{1/(\lambda\tau)}\}$. We have

$$\sup_{(c,W)\in\mathcal{W}(T,\lambda,\tau)} E[f(x;\hat{W})] - E_N[f(x,\hat{W})] \leq \sqrt{\frac{DT\log(4TN/(\tau\lambda))}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}$$

with probability $1-\rho$ through the same argument as in the case of combinatorial objective (32). Combining the above *estimation error* with the *bias* and *optimization error* in (35), we have

$$\begin{aligned}&E[f(x;W)] - E[f(x;W^*)] \\ &\leq \frac{\tau}{2}R^2 + \lambda K + \frac{2\gamma K}{\beta T} + \sqrt{\frac{2(1-\mu)K}{\mu\beta}}\lambda \\ &\quad + \sqrt{\frac{DT\log(4TN/(\tau\lambda))}{2N} + \frac{1}{2N}\log(\frac{1}{\rho})}\end{aligned}$$

Choosing $T = \frac{2\gamma K}{\beta}(\frac{1}{\epsilon})$, $\lambda = \tau = \frac{1}{\sqrt{N}}$ and $N \gtrsim \frac{DT}{\epsilon^2} = \frac{DK}{\epsilon^3}$ gives the result. □

7.5. Proof of Theorem 4

Proof. Since W^* is of rank K, we have $span(\Theta^*) = span(Z^*)$. Therefore, from condition 2,

$$span(\Theta^*) \cap \{0,1\}^N \setminus \{0\} = \{Z^*_{:,j}\}_{j=1}^K. \tag{36}$$

For any $(Z, W) : ZW = \Theta^*$, we have $Z \in span(\Theta^*)$ since $Z = \Theta^* V \Sigma^{-1} U^T$ where $U\Sigma V^T$ is the SVD of W with $\Sigma : K \times K$. Then by (36) we know that $Z = Z^*$. Then it follows $W = W^*$ since the linear system $\Theta^* = Z^* W$ has unique solution for W. □

7.6. Proof of Theorem 5

Proof. The solution of (21) satisfies

$$Z_S W_S = X = Z^* W^*.$$

Since W_S has full row-rank, we have $rank(Z_S) = rank(X) = rank(Z) = K$ by condition 1 in Theorem 4. Then let $W_S = U\Sigma V^T$ be the SVD of W_S with $\Sigma : |S| \times |S|$, we have

$$Z_S = XV\Sigma^{-1}U^T = Z^* W^* V \Sigma^{-1} U^T \in span(Z^*).$$

Then by condition 2 in Theorem 4, the columns of Z_S can only be in $\{Z^*_{:,j}\}_{j=1}^K$, which implies Z_S equal to Z^* up to a permutation. Then we know $|S| = K$ and by Theorem 4 W_S also equals W^* up to a permutation. □

7.7. Proof of Theorem 8

Proof. By an application of Theorem1 of (Negahban et al., 2009), for $\lambda \geq \|\nabla f(c^*)\|_\infty$, we have the following bound on the ℓ_2 norm of $\hat{c} - c^*$:

$$\|\hat{c} - c^*\|_2 \leq \frac{\sqrt{K^*}\lambda}{\kappa_n},$$

where $\|\nabla f(c^*)\|_\infty$ is given by:

$$\|\nabla f(c^*)\|_\infty = \max_{z \in \{0,1\}^N} \frac{1}{2N^2\tau} \|A^{*T} z\|_2^2,$$

where A^* is defined as:

$$\begin{aligned} A^* &= (I - Z_S(Z_S^T Z_S + N\tau I)^{-1} Z_S^T) X \\ &= (I - Z_S(Z_S^T Z_S + N\tau I)^{-1} Z_S^T)(Z_S W^* + \epsilon) \end{aligned} \tag{37}$$

Given $P = Z_S(Z_S^T Z_S + N\tau I)^{-1} Z_S^T$, it can be seen that A^* can be rewritten as :

$$A^* = (I - P)\epsilon + (I - P)(Z_S W^*).$$

□

7.8. ell_2 error bounds on the coefficient vector $\hat{c}$

Theorem 8. *Let c^* be the true underlying vector, with support S and sparsity K^*. Let $\hat{c}$ be the minimizer of $F(c)$, defined in Equation (7). Define the* noise-level *term*

$$\rho_n := \max_{z \in \{0,1\}^N} \frac{1}{2N^2\tau} \|A^{*T} z\|_2^2,$$

where $A^* = (I - P)\epsilon + (I - P)(Z_S W^*)$ where

$$P = Z_S(Z_S^T Z_S + N\tau I)^{-1} Z_S^T.$$

Let κ_n be the restricted strong convexity *term defined as :*

$$\kappa_n := \inf_{\Delta \in \mathcal{C}} \left\{ f(c^* + \Delta) - f(c^*) - \langle \nabla f(c^*), \Delta \rangle \right\},$$

where $\mathcal{C} = \{c \| \|c_{S^c}\|_1 \leq 3\|c_S\|_1\}$. Then, if the regularization parameter is set as $\lambda \geq \rho_n$, we have the following bound on the norm of the error $\hat{c} - c^*$:

$$\|\hat{c} - c^*\|_2 \leq \frac{\rho_n \sqrt{K^*}}{\kappa_n}.$$

7.9. Proof of Theorem 6

Proof. Note that the optimization problem in Equation (22) can be rewritten as:

$$\begin{aligned} &\underset{Z \in \{0,1\}^N}{argmin} \tfrac{1}{2N} \|E + (Z^* - Z)\|_2^2 \\ &= \tfrac{1}{2N} \sum_{i=1}^N \underset{Z_i \in \{0,1\}}{argmin} (E_i + (Z_i^* - Z_i))^2 \end{aligned} \tag{38}$$

So, we have the following closed form expression for $\hat{Z}$:

$$\hat{Z}_i = \begin{cases} 1 & \text{if } Z_i^* + E_i \geq 0.5 \\ 0 & \text{o.w} \end{cases}.$$

We now compute the probability that $Z_i^* \neq \hat{Z}_i$:

$$\begin{aligned} \mathbb{P}(Z_i^* \neq \hat{Z}_i) &= \mathbb{P}(E_i \geq 0.5) * \mathbb{P}(Z_i^* = 0) \\ &\quad + \mathbb{P}(E_i \leq -0.5) * \mathbb{P}(Z_i^* = 1) \\ &\geq \min\{\mathbb{P}(E_i \geq 0.5), \mathbb{P}(E_i \leq -0.5)\} \geq c, \end{aligned} \tag{39}$$

for some positive constant c. We now use the fact that $\mathbb{E}((Z_i^* - \hat{Z}_i)^2) = \mathbb{P}(Z_i^* \neq \hat{Z}_i)$ to complete the proof of the Lemma. □