MixLasso: Generalized Mixed Regression via Convex Atomic-Norm Regularization

Ian E.H. Yen[1,2], Wei-Cheng Lee[3], Sung-En Chang[3], Kai Zhong[4], Shou-De Lin[3] and Pradeep Ravikumar[1]

[1]Carnegie Mellon University. [2]Snap Inc. [3]National Taiwan University. [4]Amazon Inc.

Abstract

- In this work, we propose a novel convex estimator (MixedLasso) for Generalized Mixed Regression models.
- 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 GMR.
- In experiments, the MixLasso significantly outperforms other methods when there is a larger number of latent regression functions.

Generalized Mixed Regression Models

- Generalized Mixed Regression Model (GMR) is a generalization of Mixed Regression, where each response is an additive combination of latent regression functions.
- In **Generalized Mixed Regression**, each response

$$\boldsymbol{y}_i = \sum_{k=1}^{K} z_{i,k} f_k(\boldsymbol{x}_i) + \omega$$

where $\boldsymbol{y}_i \in \mathbb{R}$: response, $x_i \in \mathbb{R}^D$: explanatory variable, $z_{i,k} \in \{0,1\}, k = 1,\dots,K$ binary latent indicators, $f_k(\boldsymbol{x}_i) : \mathbb{R}^D \to \mathbb{R}$: is the regression function of k-th component, and $\omega \in \mathbb{R}$: noise.

- **Standard Mixed Regession** is a special case with $\|\boldsymbol{z}_i\|_0 = 1$.
- **Goal** is to find $\mathcal{F} := \{f_k(\boldsymbol{x})\}_{k=1}^K$ minimizing the risk

$$r(\mathcal{F}) := \mathbb{E}\left[\min_{\boldsymbol{z}\in\{0,1\}^K} \frac{1}{2}(y - \sum_{k=1}^{K} z_k f_k(\boldsymbol{x}))^2\right] \quad (1)$$

which yields a trade-off between risk and size of K.

- **We focus on** two family of functions:
  - **linear case**: $f_k(\boldsymbol{x}) := \langle \boldsymbol{w}_k, \boldsymbol{x}\rangle$.
  - **Non-linear extension**: $f_k(\boldsymbol{x})$ lying in a Reproducing Kernel Hilbert Space (RKHS) $\mathcal{H}$ with respect to some Mercer kernel $\mathcal{K}(\cdot,\cdot)$.

Related Works & Results

- **Existing Approaches:**
  - MCMC, Variational Bayes:
    No finite-time theoretical guarantee.
  - convex relaxation based on nuclear norm :
    Restricted to two components, with Gaussian assumptions on the input matrix.
  - Tensor (Spectral) Methods:
    have high sample complexity w.r.t. D or K, and isotropic Guassian assumptions on the inputs.
- **This Paper:**
  - A convex estimator — MixLasso.
  - Low-order polynomial runtime and sample complexity.
  - No restrictive assumption on the inputs.

Convex Estimation via Atomic Norm (linear function)

- Regularized Empirical Risk Minimization:

$$\min_{W\in\mathbb{R}^{K\times D}, \boldsymbol{z}_i\in\{0,1\}^K} \frac{1}{2N}\sum_{i=1}^{N}(y_i - \boldsymbol{z}_i W \boldsymbol{x}_i)^2 + \frac{\tau}{2}\|W\|_F^2.$$

- Given $Z := (\boldsymbol{z}_i)_{i=1}^N$, the dual problem w.r.t. W is:

$$\min_{M=ZZ^T\in\{0,1\}^{N\times N}} \underbrace{\left\{\max_{\alpha\in\mathbb{R}^N} \frac{-1}{\tau} tr(\mathcal{D}(\alpha)XX^T\mathcal{D}(\alpha)ZZ^T) - \sum_{i=1}^{N} L^*(y_i, -\alpha_i)\right\}}_{g(M)}.$$

- **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 MixLasso estimator:

$$\min_{M_{\mathbb{R}_+^{N\times N}}} g(M) + \lambda\|M\|_{\mathcal{S}}.$$

- Equivalently, one can solve the estimator by

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

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:

$$j^* = \underset{j}{argmax} \; -\nabla_j f(c) = \underset{z\in\{0,1\}^N}{argmax} \langle -\nabla g(M), zz^T\rangle \quad (2)$$

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

for $t = 1 \dots T$ **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:

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

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

end for.

- Finding approximate greedy coordinate costs $O(ND)$ (via SDP).
- Evaluating $\nabla g(M)$: a least-square problem of cost $O(D^3|\mathcal{A}|^3)$.
- Each iteration costs $\underbrace{O(ND)}_{\text{MAX-CUT}} + \underbrace{O(D^3|\mathcal{A}|^3)}_{\text{Least-Square}}$

Non-linear Extension

- For $f_k(\boldsymbol{x})$ lies in RKHS $\mathcal{H}$ generated by $\mathcal{K}(\cdot,\cdot)$, given $\{\boldsymbol{z}_i\}_{i=1}^N$:

$$\min_{f_k\in\mathcal{H}} \frac{1}{2N}\sum_{i=1}^{N}\left(y_i - \sum_{k=1}^{K} z_{ik} f_k(\boldsymbol{x}_i)\right)^2 + \frac{\tau}{2}\sum_{k=1}^{K}\|f_k\|_{\mathcal{H}}^2 \quad (3)$$

- *Representer Theorem* ensures an expression of the form

$$f_k^*(\boldsymbol{x}) = \sum_{i=1}^{N} \alpha_i z_{ik} \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}), \; k \in [K],$$

for the minimizers:

- Similar MixLasso estimator with

$$g(M) := \max_{\alpha\in\mathbb{R}^N} -\frac{1}{2\tau} tr(\mathcal{D}(\alpha)Q\mathcal{D}(\alpha)M) - \sum_{i=1}^{N} L^*(y_i, -\alpha_i) \quad (4)$$

where $Q : N \times N$ is the kernel matrix with $Q_{ij} = \mathcal{K}(\boldsymbol{x}_i, \boldsymbol{x}_j)$.

Theoretical Results: Risk Bound

Let W^* be the minimizer of risk (1) with K components and $||W^*||_F \leq R$ and $\hat{W} = D(\sqrt{\boldsymbol{c}_{\mathcal{A}}})W$, then $\hat{W}$ with probability $1-\rho$ satisfies

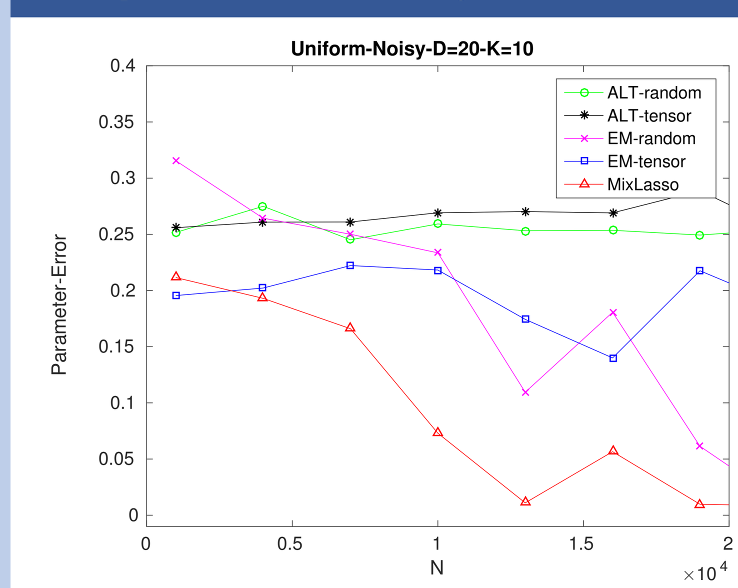
$$r(\hat{W}) \leq r(W^*) + \epsilon$$

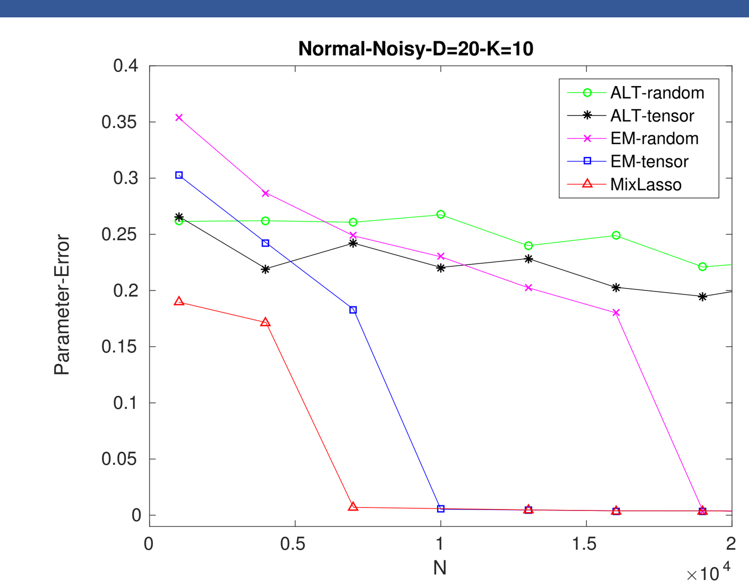
as long as

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

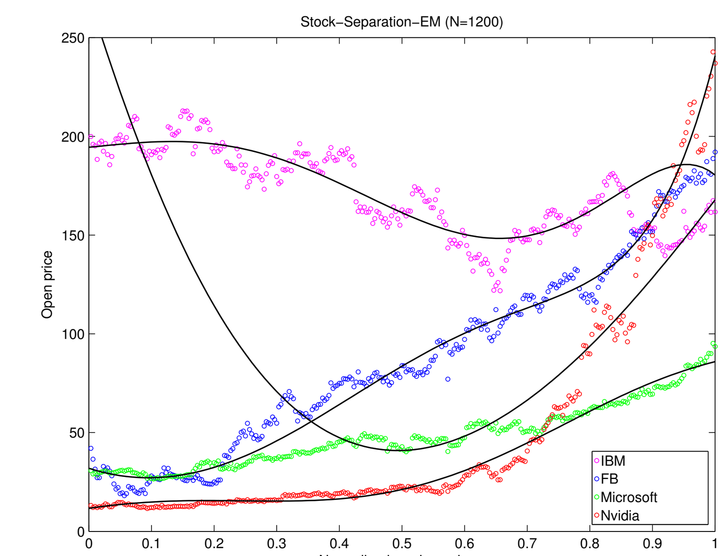
- The number of output components $\hat{K} = O(K/\epsilon)$
- The result trades between risk and sparsity.
- No assumption on W except that of boundedness.
- The sample complexity is (quasi) linear to D and K.

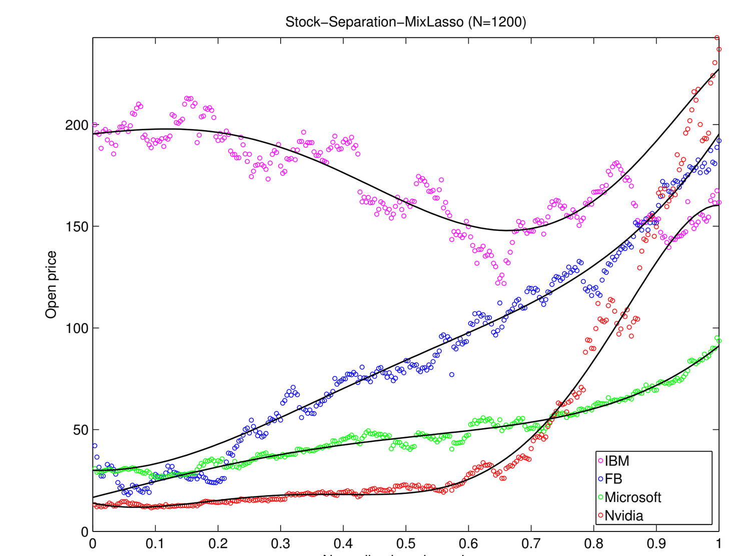
Experiments on Synthetic data





Experiments on Stock Data





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