Latent Sparse Modeling of Longitudinal Multi-dimensional Data

Ko-Shin Chen1 Tingyang Xu2 Jinbo Bi1
1Department of Computer Science and Engineering, University of Connecticut, Storrs, CT, USA
2Tencent AI Lab, Shenzhen, China

Introduction
In this work, we propose a tensor-based model where outcomes not only depend on the current observation but also from multiple previous consecutive observations. Simultaneously, the model determines the temporal contingency and the most influential features along each dimension of the tensor data.

K-way Tensor & latent L_{F,1} Regularizer
- We decompose a K-way tensor $W$ as $W = \sum_{k=1}^{K} W_k$ such that $W_k$ is sparse along the k-th direction.
- Example ($K = 3$)
- To select the most important time points and features, we introduce the latent L_{F,1} regularizer:
  $$ R(\Phi) = \sum_{k=1}^{K} \lambda_k \left( \sum_{j=1}^{d_k} \left\| (W_k)_{(j)(k)} \right\|_F \right) $$
  Here $\Phi = (W_1, \ldots, W_K)$ and $\| \cdot \|_F$ denotes the Frobenius norm of a tensor.

Model Formulation
- Assume the outcome $y_t$ depends on current and previous $\tau$ observations: $x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-\tau}$. The covariate $X(t) = [x_t, x_{t-1}, x_{t-2}, \ldots, x_{t-\tau}]$ is a K-way tensor.
- We apply the quadratic inference function (QIF) to deal with within-sample correlation so that the correlation structure does not need to be pre-specified.

$$ \min_{\Phi=(W_1,\ldots,W_K)} F(\Phi) := Q(X, y - \mathbb{E}[y], W) + R(\Phi) $$
  such that $W = \sum_{k=1}^{K} W_k$.

Algorithm: Linearized Block Coordinate Descent
Given $W_1^{(r)}, \ldots, W_K^{(r)}$, for each fixed $k$, we obtain the updated $W_k^{(r+1)}$ by solving

$$ \min_{W_k} \frac{1}{2} \left\| W_k - (W_k^{(r)} - C H (Q^{(r)})) \right\|_F^2 + C \lambda_k \left( \sum_{j=1}^{d_k} \left\| (W_k)_{(j)(k)} \right\|_F \right) $$

Theorems
- **Theorem 1**: Under certain assumptions, when the sample size $m \to \infty$, we have $\hat{W} \to W^*$ in probability and $\sqrt{m} \text{vect}(\hat{W} - W^*) \to \mathcal{N}(0, \Sigma)$ in distribution for some $\Sigma$.
- **Theorem 2**: If the initial tuple $\Phi^{(0)}$ is within a convex neighborhood of a minimizer $\hat{\Phi}$, the algorithm converges and $F(\Phi^{(r)}) - F(\hat{\Phi}) \leq \frac{\|W_k^{(0)} - \hat{W}\|^2}{2 \tau r C}$ for all $r \geq 1$.

Feature Selection
- **fMRI Data**
  - Goal: to predict MMSE score.
  - Direction 1: brain areas (67).
  - Direction 2: brain properties (4).
    - CV – Cortical Volume
    - SA – Surface Area
    - TA – Thickness Average
    - TS – Thickness SD
  - Direction 3: time (month).
- **EEG Data**
  - Goal: to predict responses (±1) on Sternberg task.
  - Direction 1: processing stages
    - (baseline, encoding, retention, retrieval)
  - Direction 2: electrodes (12).
  - Direction 3: frequency components ($\delta, \theta, \alpha, \beta, \gamma$).
  - Features selected for schizophrenia patients and healthy control participants.

Conclusion
1. The tensorQIF model finds true coefficient.
2. The corresponding optimization problem can be solved by an algorithm with a sublinear convergence rate.
3. The proposed method enabled interpretation and summary across all dimensions.

Acknowledgements: This work was partially supported by NSF grants DMI-1356655, CCF-1514357, IIS-1718738, as well as NIH grants R01DA037349 and K02DA043063 to Jinbo Bi.