Generalized Block-Diagonal Structure Pursuit
Learning Soft Latent Task Assignment against Negative Transfer

Zhiyong Yang, Qianqian Xu, Yangbangyan Jiang, Xiaochun Cao, and Qingming Huang

IIE, CAS; ICT, CAS; UCAS; BDKM, CAS;PCL.

Overview

To avoid negative transfer, we propose a novel MTL method with the following contribution:

- ✓ disentangled latent task assignments with a block-diagonal constraint
- ✓ a novel regularizer for generalized block-diagonal structure pursuit
- ✓ explicit connection with optimal transport
- ✓ theoretical guarantee for structual recovery.

Latent Task Representation: A Probabilist View

- ➥ Given T tasks, the per-task parameter W is defined as $W = [W^{(1)}, \cdots, W^{(T)}] \in \mathbb{R}^{d \times T}$.
- ➥ To model the relationship among the tasks $o_1, \cdots, o_T$, we assume that $W^{(i)}$ could be represented as a linear combination of latent tasks $l_1, \cdots, l_k$, with $W = LS$.
- ➥ From a probabilist perspective, we regard $S_{i,j}$ as $\mathbb{P}(l = i | o = j)$, namely the possibility of choosing l_i to represent o_j.
- ➥ Another variable of interest is the joint distribution matrix $S^{\ddagger}{}_{i,j} = \mathbb{P}(l = i, o = j)$. Given marginal distribution **a**, **b** for l, $ø$, we must have $S^{\ddagger} 1_T = \mathbf{a}$, $S^{\ddagger} 1_k = \mathbf{b}$.
- ➥ In order to avoid suppress, we hope the possibility to assign l_i to o_j is nonzero if and only if (i, j) belongs to the same group.This leads to a block-diagonal structure of $S^{\ddagger}$ up to row and column permutations.

Block-diagonal Structure Pursuit

Auxiliary Bipartite Graph. $A_{l \cup o} = \begin{bmatrix} 0 & S^{\ddagger} \\ S^{\ddagger\top} & 0 \end{bmatrix}$, $\Delta(S^{\ddagger}) = diag(A_{l \cup o} 1) - A_{l \cup o}$. The following shows that squeezing the bottom K eigenvalues of $\Delta(S^{\ddagger})$ Then leverages the block diagonal property.

Theorem

If $S^{\ddagger} \in \Pi(\mathbf{a}, \mathbf{b})$, $\chi_S = K$ holds if and only if $\dim(Null(\Delta(S^{\ddagger}))) = K$, i.e, the 0 eigenvalue of $\Delta(S^{\ddagger})$ has multiplicity K. Moreover, denote $\mathcal{A}^{(i)}$ as the set of latent and output tasks belonging to the i-th block of S, the eigenspace of 0 is spanned by $\iota_{\mathcal{A}^{(1)}}, \iota_{\mathcal{A}^{(2)}}, \cdots, \iota_{\mathcal{A}^{(K)}}$, where $\iota_{\mathcal{A}^{(i)}} \in \mathbb{R}^{(k+T)\times 1}$, $[\iota_{\mathcal{A}^{(i)}}]_j = 1$ if $j \in \mathcal{A}^{(i)}$, otherwise $[\iota_{\mathcal{A}^{(i)}}]_j = 0$.

From the variational property of eigenvalues, we reach the regularizer $\Omega(S^{\ddagger}) = \inf \left\{ \left\langle \Delta(S^{\ddagger}), U \right\rangle : U \in \mathcal{M} \right\}$, where

$$\mathcal{M} = \{U : U \in \mathbb{S}^N, I \succeq U \succeq 0, tr(U) = K\}$$

.

Objective Function

Exact problem: Loss + Reg. of L + Structural Reg.

$$\min \tilde{\mathcal{J}} + \Omega_1 + \Omega_2 \ \ s.t. \ S^{\ddagger} \in \Pi(\mathbf{a}, \mathbf{b}), \ U \in \mathcal{M}, \ S = TS^{\ddagger}. \quad (1)$$

Inexact problem: Exact + Variable Splitting Between S and $TS^{\ddagger}$.

$$\min \tilde{\mathcal{J}} + \Omega_1 + \Omega_2 + \Omega_3 \ \ s.t. \ S^{\ddagger} \in \Pi(\mathbf{a}, \mathbf{b}), \ U \in \mathcal{M}. \quad (2)$$

$\Omega_1 = \alpha_1 \cdot ||L||_F^2/2$, $\Omega_2 = \alpha_3 \cdot \langle \Delta(S^{\ddagger}), U \rangle$, $\Omega_3 = \alpha_2 \cdot d(S, TS^{\ddagger})/2$

Optimization

We present an alternative optimization method to solve.

L, S **subproblem.** strongly convex, could be solved from off-the-shelf tools.

U **subproblem.** $U = V_K V_K^{\top}$, where V_K denotes eigenvectors associated with the smallest k eigenvalues of $\Delta(S^{\ddagger})$. Define f_i from $V_K = [f_1, \cdots, f_{k+T}]^{\top}$. f_i has a strong grouping power when $\chi_{S^{\ddagger}} = K$.

$S^{\ddagger}$ **subroutine**: With U updated with $U = V_K V_K^{\top}$, we reformulate the subproblem as a regularized optimal transport problem:

Proposition (Regularized OT Reformulation)

Reformulation. The $S^{\ddagger}$ subproblem could be reformulated as:

$$\min_{S^{\ddagger} \in \Pi(\mathbf{a},\mathbf{b})} \frac{\vartheta}{2} \|S^{\ddagger} - \bar{S}\|_F^2 + \langle \mathcal{D}, S^{\ddagger} \rangle \qquad (Primal)$$

Connection with OT. Under mild conditions, we have: $0 \le \mathcal{J}_{REG} - \mathcal{J}_{OT} = O(\vartheta/T)$.

Dual Solution The dual problem of (Primal) could be solved from:

$$\underset{h,\, g}{\operatorname{argmin}} \ \frac{1}{2\vartheta} \cdot \|(h \oplus g - \mathcal{D} + \vartheta \bar{S})_+\|_F^2 - \langle h, \mathbf{a} \rangle - \langle g, \mathbf{b} \rangle,$$

with $S^{\ddagger\star} = \left[\frac{h^\star \oplus g^\star - \mathcal{D}}{\vartheta} + \bar{S} \right]_+$.

Negative transfers are punished via a large transportation cost, while positive transfers within group is encouraged with an almost zero cost.

Theoretical Analysis

The hypothesis space $\mathcal{H}$:

$$\left\{ \left\{ \hat{Y}^{(i)}(X_i^{(t)}) = (LS^{(i)})^{\top} X_i^{(t)} \right\}_{ti} : \|L\|_F^2 \le \xi_1, \right.$$
$$\left. d(S, TS^{\ddagger}) \le \xi_2, \ \left\langle \Delta(S^{\ddagger}), U \right\rangle \le \xi_3, S^{\ddagger} \in \Pi(\mathbf{a}, \mathbf{b}), U \in \mathcal{M} \right\}$$

$\xi_1 = 2\mathcal{J}_0/\alpha_1, \xi_2 = 2\mathcal{J}_0/\alpha_2, \xi_3 = 2\mathcal{J}_0/\alpha_3$

Theorem (Performance Bounds)

Generalization Bound. Picking $\xi_1 = O(1/T)$, $\xi_2 = O(1/T)$, if ℓ is Lipschitz continuous, we have: $|\mathcal{R}(L, S) - \hat{\mathcal{R}}(L, S)| = O_P((nT)^{-1/2})$.

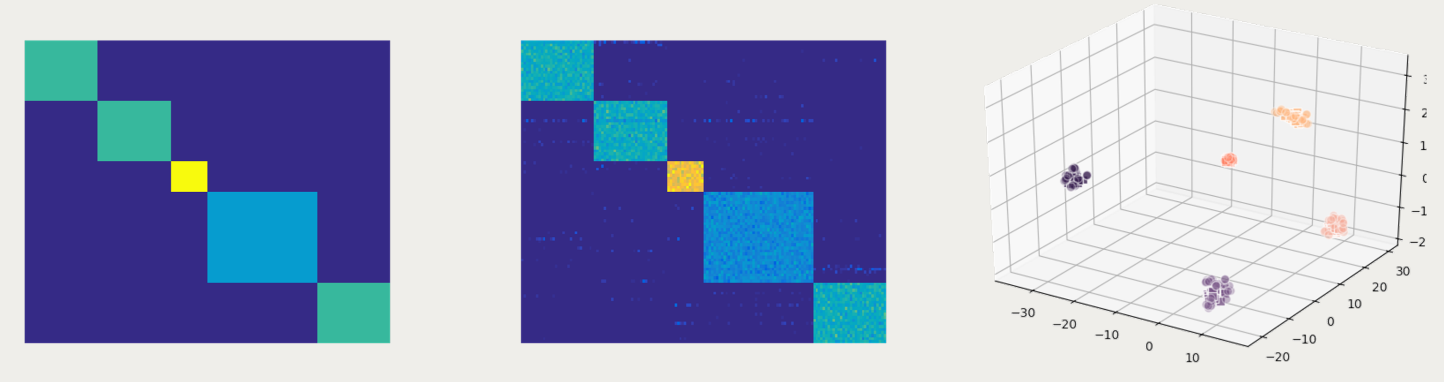
Implicit Spectrum Bound. Picking $\xi_3 = (1/T)^{-3/2}$, $\xi_2 = (1/T^2)$, we have $\sum_{i=N-K+1}^{N} \lambda_i(\Delta(S)) = O(T^{-1/2})$.

Structure Recovery Bound. Assume that $k \le T$, for all $S^{\ddagger}$ obtained from the space $\mathcal{H}$ such that $\lambda_{K+1}(\Delta(S^{\ddagger})) > \lambda_K(\Delta(S^{\ddagger})) > 0$, there is a co-partition, such that:

$$\|S^{\ddagger supp^c}\|_1 = O\left(k^{1/2}/(T \cdot \lambda_{K+1}(\Delta(S^{\ddagger}))\right),$$
$$\frac{|supp^c|}{kT} = O\left(k^{-1/2}/\left(T \cdot \lambda_{K+1}(\Delta(S^{\ddagger}))\right)\right).$$

Simulated Dataset

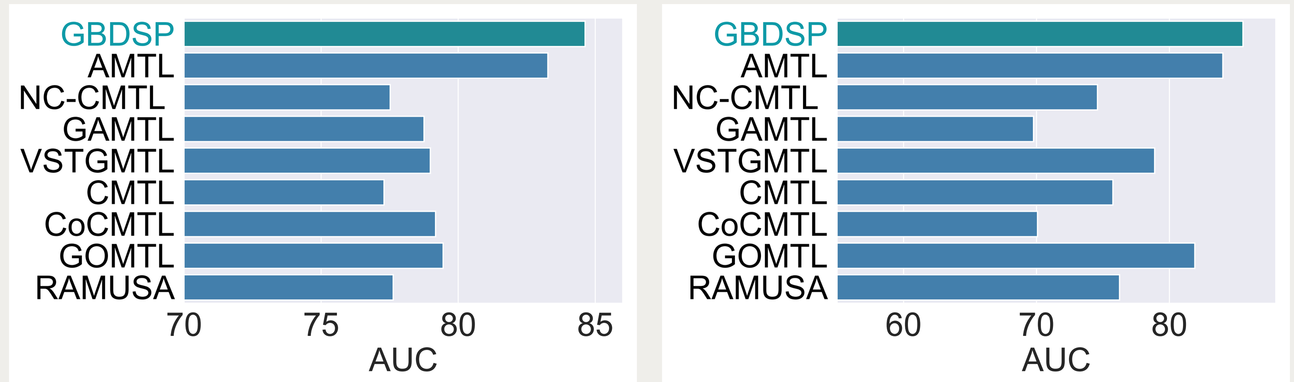
To test the effectiveness of GBDSP we generate a simple simulated annotation dataset with $T = 150$ simulated tasks, where the dataset is produced according to the assumption in our model.



(a) (b) (c)

Figure: Visualizations over the Simulated Dataset. (a) shows The true LATM; (b) shows the LATM recovered by GBDSP(c) shows the spectral embedding GBDSP.

Real World Datasets



(a) (b)

Figure: (a) Performance Comparison Curve over the Simulation Dataset, with varying training data ratio. (b) Performance Comparison over the Sun Dataset.

Table: Performance Comparison over General MTL Datasets (mean $\pm$ std)

Algorithms	AWA2-Attr (↑)	AWA2-Cls (↑)	School (↓)
RAMUSA	88.60±0.66	93.06±0.53	10.52±0.09
GOMTL	89.56±0.33	88.22±1.18	10.26±0.11
CoCMTL	92.29±0.35	94.69±0.73	12.06±0.09
CMTL	92.95±0.35	94.81±0.70	12.06±0.09
VSTGMTL	89.31±0.39	92.03±0.94	10.17±0.08
GAMTL	89.39±0.42	92.55±0.53	10.50±0.12
NC-CMTL	92.99±0.32	95.10±0.60	10.53±0.12
AMTL	92.15±0.34	95.76±0.44	12.15±0.09
GBDSP	92.73±0.29	97.86±0.22	10.10±0.08

Contact Information

Our Group: https://qmhuang-ucas.github.io/

Email me : yangzhiyong@iie.ac.cn

Visit my home page from QR code