Unveiling Groups of Related Tasks in Multi-Task Learning

Jordan Frecon[1], Saverio Salzo[1], Massimiliano Pontil[1,2]

[1] Computational Statistics and Machine Learning, Istituto Italiano di Tecnologia, Genova, Italy
[2] Department of Computer Science, University College London, London, UK

ICPR 2020, Milano, Italy

1. Introduction

A common approach in multi-task learning is to encourage the tasks to share a low dimensional representation by using trace norm regularization. In this paper, we extend this approach by allowing the tasks to partition into different groups, within which trace norm regularization is separately applied. We propose a smooth continuous bilevel optimization framework to simultaneously identify groups of related tasks and learn a low dimensional representation within each group.

2. Groupwise Trace Norm Regularization

Goal: Given a dataset $\{y_t^{\mathrm{trn}}, X_t^{\mathrm{trn}}\}_{t=1}^T$ assumed organized into L groups of related tasks $\{\mathcal{G}_1, \dots, \mathcal{G}_L\}$, find a regressor W such that

- (linear regression) for every $t \in \{1, \dots, T\}$, $y_t \approx X_t w_t$
- (low dimension) for every $l \in \{1, \dots, L\}$, the restriction $W_{\mathcal{G}_l}$ is low rank

Optimization problem: Given $\{y_t^{\mathrm{trn}}, X_t^{\mathrm{trn}}\}_{t=1}^T$ and a partition $\mathcal{G} = \{\mathcal{G}_1, \dots, \mathcal{G}_L\}$ of the T tasks in L groups, find

$$\hat{W} \in \operatorname*{argmin}_{W \in \mathbb{R}^{P \times T}} \sum_{t=1}^T \frac{1}{2}\|y_t - X_t w_t\|^2 + \lambda \sum_{l=1}^L \|W_{\mathcal{G}_l}\|_{\mathrm{tr}},$$

for some regularization parameter $\lambda > 0$

Issue: In many applications, $\mathcal{G}$ might not be known *a priori*. Finding $\mathcal{G}$ through an exhaustive search is a challenging combinatorial problem since there are $(L^T/L!)$ possible partitions

3. Proposed Setting

Parametrization of the Groups: Let $\theta = [\theta_1 \cdots \theta_L] \in [0,1]^{T \times L}$ be the hyperparameter matrix encoding at most L groups, meaning that $\theta_{t,l} = 1$ if the t-th task belongs to $\mathcal{G}_l$, and 0 otherwise

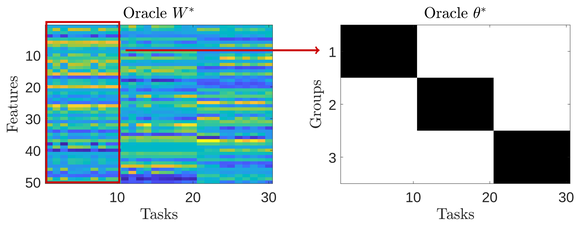


Figure: The oracle parameter matrix W^* is made of 3 groups of related tasks : $1 \to 10$, $1 \to 20$ and $21 \to 30$.

Data: Training sets $\{y_t^{\mathrm{trn}}, X_t^{\mathrm{trn}}\}_{t=1}^T$ and validation sets $\{y_t^{\mathrm{val}}, X_t^{\mathrm{val}}\}_{t=1}^T$ each one sample from a groupwise low-rank linear model $y_t = X_t w_t^* + \epsilon_t$, where $\epsilon_t \sim \mathcal{N}(0_P, \sigma^2 \mathbb{1}_P)$

Main Goal

Estimate θ^* from $\{y_t^{\mathrm{trn}}, X_t^{\mathrm{trn}}\}_{t=1}^T$ and $\{y_t^{\mathrm{val}}, X_t^{\mathrm{val}}\}_{t=1}^T$

4. Exact Bilevel Problem

Upper-level Problem:

$$\underset{[\theta_1 \cdots \theta_L] \in \Theta}{\text{minimize}}\ \mathcal{U}(\theta) := \sum_{t=1}^T \frac{1}{2}\|y_t^{\mathrm{val}} - X_t^{\mathrm{val}} \hat{w}_t(\theta)\|^2$$

where $\hat{W}(\theta) = \left[\hat{w}_1(\theta) \cdots \hat{w}_T(\theta)\right]$ solves

Lower-level Problem:

$$\underset{W \in \mathbb{R}^{P \times T}}{\text{minimize}}\ \mathcal{L}(W, \theta) := \underbrace{\left(\sum_{t=1}^T \frac{1}{2}\|y_t^{\mathrm{trn}} - X_t^{\mathrm{trn}} w_t\|^2 + \frac{\epsilon}{2}\|w_t\|^2\right)}_{f(W)\ \text{smooth}} + \underbrace{\lambda \sum_{l=1}^L \|\theta_l \odot W\|_{\mathrm{tr}}}_{g(A_\theta W)\ \text{nonsmooth}}$$

Difficulties:

- $\mathcal{L}$ is nonsmooth (since g is nonsmooth)
- $\mathcal{U}$ is nonsmooth (since $\hat{W}$ is nonsmooth)
- $\hat{W}(\theta)$ is not available in closed form

5. Approximate Bilevel Problem

Upper-level Problem:

$$\underset{[\theta_1 \cdots \theta_L] \in \Theta}{\text{minimize}}\ \mathcal{U}_K(\theta) := \sum_{t=1}^T \frac{1}{2}\|y_t^{\mathrm{val}} - X_t^{\mathrm{val}} w_t^{(K)}(\theta)\|^2$$

where $W^{(K)}(\theta) = \left[w_1^{(K)}(\theta) \cdots w_T^{(K)}(\theta)\right]$ is given by

Dual Algorithm:

$U^{(0)}(\theta)$ arbitrarily chosen
for $k = 0, 1, \dots, K-1$
$\quad U^{(k+1)}(\theta) = \mathcal{A}(U^{(k)}(\theta), \theta)$ *dual update*
$W^{(K)}(\theta) = \mathcal{B}(U^{(K)}(\theta), \theta)$ *primal dual relationship*

Goals: Find $\mathcal{A}, \mathcal{B}$ such that

- $\mathcal{U}_K$ is smooth
- $W^{(K)}(\theta) \to \hat{W}(\theta)$
- $\min \mathcal{U}_K \to \min \mathcal{U}$ and $\operatorname{argmin} \mathcal{U}_K \to \operatorname{argmin} \mathcal{U}$

When does the Approximate Bilevel Problem Converge to the Exact One?

Theorem 1: Suppose that Θ is a compact nonempty subset of $\mathbb{R}_+^{T \times L}$. If the iterates $\{W^{(K)}(\theta)\}_{K \in \mathbb{N}}$ converge to $\hat{W}(\theta)$ uniformly on Θ as $K \to +\infty$, then

$$\inf_{\theta \in \Theta} \mathcal{U}_K(\theta) \underset{K \to +\infty}{\longrightarrow} \inf_{\theta \in \Theta} \mathcal{U}(\theta) \quad \text{and} \quad \operatorname*{argmin}_{\theta \in \Theta} \mathcal{U}_K(\theta) \underset{K \to +\infty}{\longrightarrow} \operatorname*{argmin}_{\theta \in \Theta} \mathcal{U}(\theta)$$

6. Algorithmic Solution

We implement a forward-backward algorithm with Bregman distance [1, 2] for solving the dual problem:

$$\underset{U \in (\mathbb{R}^{P \times T})^L}{\text{minimize}}\ \underbrace{f^*(-A_\theta^\top U)}_{\text{smooth}} + \underbrace{g^*(U)}_{\text{nonsmooth}}$$

Mapping $\mathcal{A}$: *smooth for specific Legendre function Φ*

$$U^{(k+1)}(\theta) = \operatorname{prox}_{\gamma g^*}^{\Phi}\left(\nabla\Phi(U^{(k)}(\theta)) + \gamma A_\theta \nabla f^*(-A_\theta^\top U^{(k)}(\theta))\right)$$
$$= \underset{U \in (\mathbb{R}^{P \times T})^L}{\operatorname{argmin}}\ \gamma g^*(U) + \Phi(U) - \left\langle U, \nabla\Phi(U^{(k)}(\theta)) + \gamma A_\theta \nabla f^*(-A_\theta^\top U^{(k)}(\theta))\right\rangle$$

Mapping $\mathcal{B}$: $W^{(K)}(\theta) = \nabla f^*(-A_\theta^\top U^{(K)}(\theta))$ *smooth* ✓

Uniform Convergence

Theorem 2: For every $\theta \in \Theta$, $\|w^{(K)}(\theta) - \hat{w}(\theta)\|^2 \leq \dfrac{\mathrm{Const}\lambda}{\epsilon K}$

Overall algorithm: If $\mathcal{A}$ and $\mathcal{B}$ are smooth, then so does $\mathcal{U}_K$. Hence, one can implement the following algorithm

$$(\forall q \in \{0, \dots, Q-1\}), \quad \theta^{(q+1)} = \mathcal{P}_\Theta\left(\theta^{(q)} - \nu \nabla \mathcal{U}_K(\theta^{(q)})\right)$$

The key novelty of our approach is to devise an efficient algorithm to compute $\nabla \mathcal{U}_K(\theta^{(q)})$ by exploiting recent results on the derivative of generalized matrix functions [3].

7. Choice of Legendre function

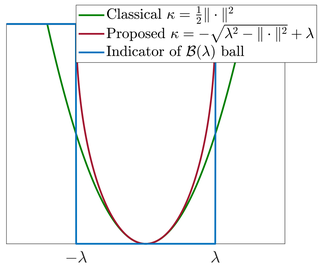
Separable Legendre function Since g is the sum of L trace norms, then g^* is separable and equal to the indicator function of $\mathcal{B}_{\mathrm{sp}}(\lambda)^L$, where $\mathcal{B}_{\mathrm{sp}}(\lambda)$ is the spectral ball of $\mathbb{R}^{P \times T}$ with radius λ. Hence, we look for a function Φ separable as well,

$$\Phi \colon V \mapsto \sum_{l=1}^L \phi(V_l) \quad \text{and} \quad \operatorname{prox}_{\gamma \iota_{\mathcal{B}_{\mathrm{sp}}(\lambda)^L}}^{\Phi} \colon V \mapsto \left(\operatorname{prox}_{\gamma \iota_{\mathcal{B}_{\mathrm{sp}}(\lambda)}}^{\phi}(V_l)\right)_{l \in \{1, \dots, L\}}$$

In order to find a smooth proximity function, we look for a Legendre function ϕ such that $\operatorname{dom} \phi = \mathcal{B}_{\mathrm{sp}}(\lambda)$.

Legendre function acting on the singular values: Given the singular value decomposition $V_l = A\operatorname{diag}[\sigma]B^\top$ with $\sigma = (\sigma_1, \dots, \sigma_r)$, we define

$$\phi \colon V_l \mapsto \sum_{i=1}^r \kappa(\sigma_i) \quad \text{and} \quad \operatorname{prox}_{\gamma \iota_{\mathcal{B}_{\mathrm{sp}}(\lambda)}}^{\phi} \colon V_l \mapsto A\operatorname{diag}[\kappa'(\sigma_i)]B^\top$$



8. Synthetic Experiments

Setting: $T = 30$ tasks arranged in $L = 3$ groups of 10 tasks each. For each task, we have $N = 10$ noisy observations ($\sigma^2 = 0.1$) and $P = 20$ features. We do not assume that the true number of groups is known. Instead, we let the methods find at most 6 groups. The regularization parameter λ is selected on a grid to minimize the validation error

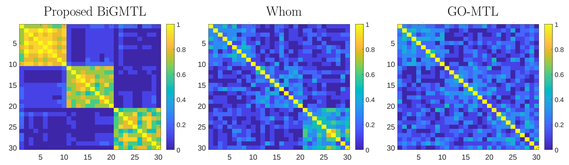


Figure: Mean group covariance matrix $\theta^\top\theta$ on the synthetic experiment. Only the proposed method manages to clearly estimate the three groups of tasks.

9. Real Data Experiments

	Animals (accuracy)	Parkinson (mse $\times 10^{-1}$)
Proposed BiGMTL	**84.96% (± 0.90)**	**4.98** (±0.20)
STL [4]	83.55% (± 0.79)	5.01 (±0.15)
Whom [5]	84.72% (± 0.65)	5.03 (±0.16)
STL2 [6]	84.06% (± 0.82)	5.11 (±0.12)
RMTL [7]	74.04% (± 4.28)	5.08 (±0.19)
GO-MTL [8]	80.51% (± 0.73)	5.95 (±0.17)
MeTaG [9]	84.35% (± 0.80)	5.03 (±0.18)

Table: Results on benchmark data sets. We report the average over multiple splits and the standard deviation in parenthesis.

Experimental results indicate the advantage of working with a variable number of groups over standard trace norm regularization (STL, STL2) and previous state-of-the-art approaches.

References

[1] H. H. Bauschke, J. Bolte, and M. Teboulle. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Mathematics of Operations Research*, 42(2):330–348, 2016.

[2] Q. Van Nguyen. Forward-backward splitting with Bregman distances. *Vietnam Journal of Mathematics*, 45(3):519–539, 2017.

[3] V. Noferini. A formula for the Fréchet derivative of a generalized matrix function. *SIAM Journal on Matrix Analysis and Applications*, 38(2):434–457, 2017.

[4] Ting Kei Pong, Paul Tseng, Shuiwang Ji, and Jieping Ye. Trace norm regularization: Reformulations, algorithms, and multi-task learning. *SIAM J. on Optimization*, 20(6):3465–3489, December 2010.

[5] Z. Kang, K. Grauman, and F. Sha. Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pages 521–528, USA, 2011. Omnipress.

[6] Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, Dec 2008.

[7] Jianhui Chen, Jiayu Zhou, and Jieping Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50, 2011.

[8] Abhishek Kumar and Hal Daume III. Learning task grouping and overlap in multi-task learning. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

[9] Lei Han and Yu Zhang. Learning multi-level task groups in multi-task learning. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.