Variable Selection and Task Grouping for Multi-task Learning

JunYong Jeong and Chi-Hyuck Jun,

Industrial and Management Engineering, POSTECH, Rep of Korea

june0227@postech.ac.kr and chjun@postech.ac.kr

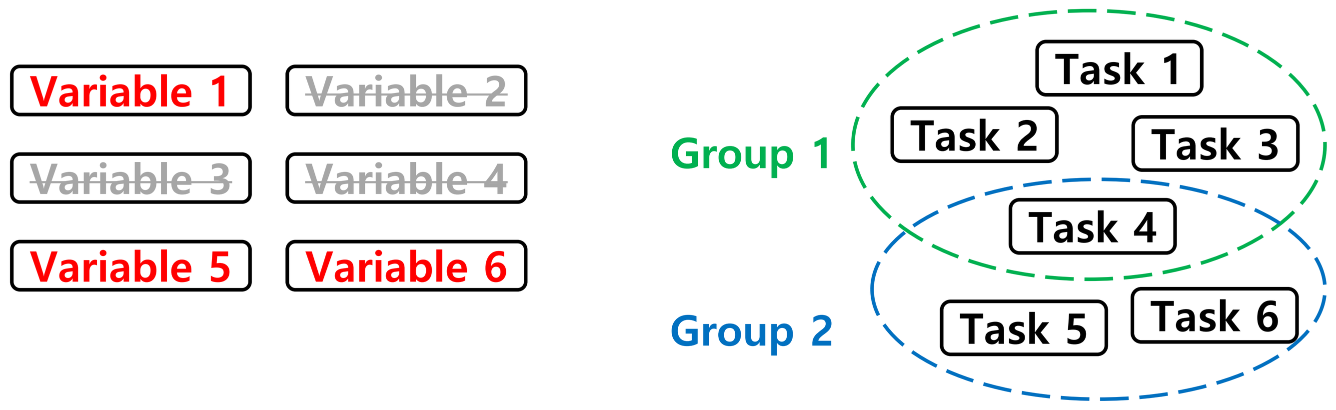
Problem and Purpose

Multi-task learning (MTL)

- D variables and T tasks

$$\mathbf{X}_j = \left[\left(\mathbf{x}_j^1\right)^T, \dots, \left(\mathbf{x}_j^{N_j}\right)^T\right]^T \in \mathbb{R}^{N_j \times D} \;\&\; \mathbf{y}_j = \left[y_j^1, \dots, y_j^{N_j}\right]^T \in \mathbb{R}^{N_j}, \quad t = 1, \dots, T$$

Variable Selection and Task Grouping MTL (VSTG-MTL)



Formulation

Linear model

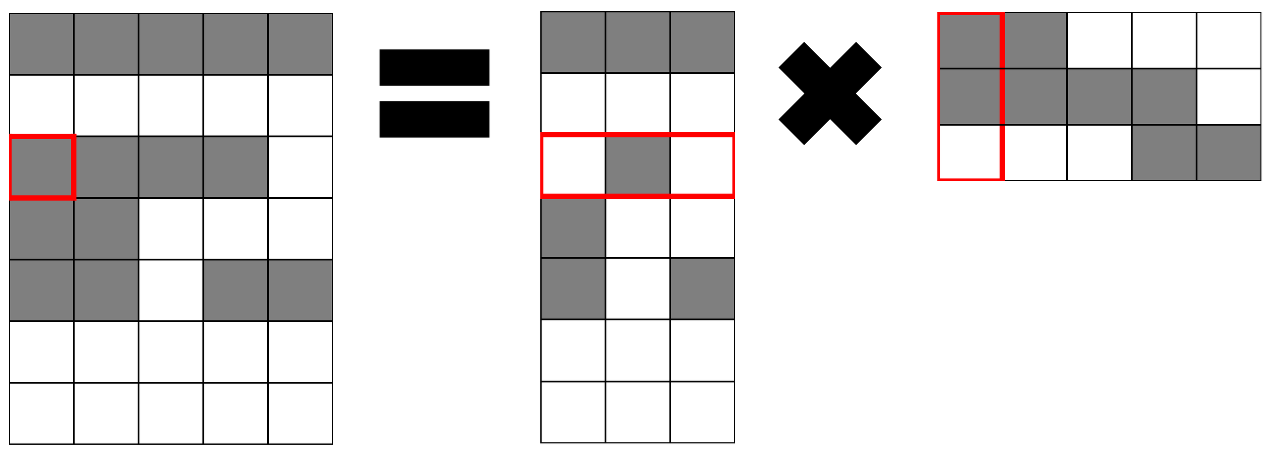
$$\hat{y}_j^n = \begin{cases} \mathbf{w}_j^T \mathbf{x}_j^n & y_j^n \in \mathbb{R} \\ \dfrac{1}{1+\exp\left(-\mathbf{w}_j^T \mathbf{x}_j^n\right)} & y_j^n \in \{-1,1\} \end{cases}$$

$\mathbf{W} = (w_{ij}) = [\mathbf{w}_1, \dots, \mathbf{w}_T] \in \mathbb{R}^{D \times T}$: Coefficient matrix

Main idea: Low-rank factorization & Sparsity

$$\mathbf{W} = \mathbf{U}\mathbf{V}$$

where $\mathbf{U} \in \mathbb{R}^{D \times M}$ is the variable-latent matrix, $\mathbf{V} \in \mathbb{R}^{M \times T}$ is the latent-task matrix, and $M \ll \min(D, T)$ is the number of latent basis



Importance vector

$$w_{ij} = \mathbf{u}^i \mathbf{v}_j$$

⇒ $\mathbf{u}^i \in \mathbb{R}^{1 \times M}$: ith row vector & importance vector of ith variable

⇒ $\mathbf{v}_j \in \mathbb{R}^M$: jth column vector & importance vector for jth task

Linear combination of latent basis vectors

$$\mathbf{w}_j = \mathbf{U}\mathbf{v}_j = \sum_{m=1}^{M} v_{mj} \mathbf{u}_m$$

⇒ $\mathbf{u}_m \in \mathbb{R}^D$: mth column vector & mth latent basis vector

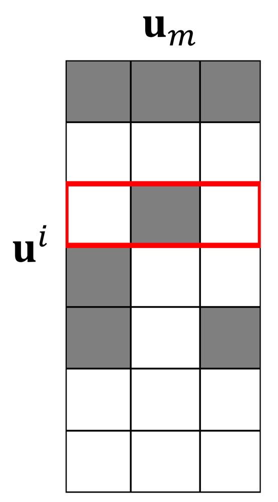
⇒ $\mathbf{v}_j$: weighting vector in the linear combination for jth task

Representation Learning

$$y_j^n = \mathbf{v}_j^T \mathbf{U}\mathbf{x}_j^n = \mathbf{v}_j^T\left(\mathbf{U}\mathbf{x}_j^n\right)$$

⇒ $\mathbf{U}\mathbf{x}_j^n \in \mathbb{R}^M$: new representation where correlation would exist

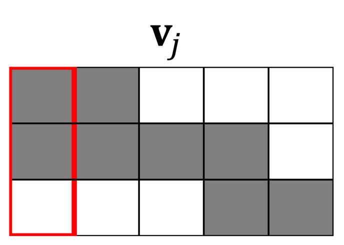
⇒ $\mathbf{v}_j \in \mathbb{R}^M$: coefficient vector on the new representation for jth task



Variable-latent matrix U

- Sparsities between and within the rows, variable importance vectors $\mathbf{u}^i$

⇒ Flexible variable selection



Latent-task matrix V

- Sparsity within the column, task weighting vector $\mathbf{v}_j$

⇒ Task grouping

Optimization

Penalized problem

$$\min_{\mathbf{U},\mathbf{V}} \sum_{j=1}^{T} \frac{1}{N_j} L(\mathbf{y}_j, \mathbf{X}_j \mathbf{U}\mathbf{v}_j)$$

$s.t$

C1: $\|\mathbf{U}\|_1 = \sum_{i=1}^{D} \|\mathbf{u}^i\|_1 \le \alpha_1$,

C2: $\|\mathbf{U}\|_{1,\infty} = \sum_{i=1}^{D} \|\mathbf{u}^i\|_\infty \le \alpha_2$,

C3: $\sum_{j=1}^{T}\left(\|\mathbf{v}_j\|_k^{sp}\right)^2 \le \beta$

C1 & C2: L1,1 & L1,inf norm

⇒ impose sparsities between and within the variable importance vectors $\mathbf{u}^i$

⇒ perform variable selection

C3: Squared k-support norm (Argyriou *et al.*, 2012)

⇒ impose sparsity within the task weighting vector $\mathbf{v}_j$ & consider possible correlation

⇒ perform task grouping

Transformation to a regularized problem

$$\min_{\mathbf{U},\mathbf{V}} \sum_{j=1}^{T} \frac{1}{N_j} L(\mathbf{y}_j, \mathbf{X}_j \mathbf{U}\mathbf{v}_j) + \gamma_1 \|\mathbf{U}\|_1 + \gamma_2 \|\mathbf{U}\|_{1,\infty} + \mu \sum_{j=1}^{T}\left(\|\mathbf{v}_j\|_k^{sp}\right)^2$$

Alternating Optimization

1. Learn a ridge regression for each task to compute initial coefficient vector

$$\mathbf{w}_j^{init} := \operatorname*{argmin}_{\mathbf{w}} \frac{1}{N_j} L(\mathbf{y}_j, \mathbf{X}_j \mathbf{w}) + \sqrt{\gamma_1^2 + \gamma_2^2 + \mu^2}\, \|\mathbf{w}\|_2^2$$

$$\mathbf{W}^{init} := \left[\mathbf{w}_1^{init}, \dots, \mathbf{w}_T^{init}\right] \in \mathbb{R}^{D \times T}$$

2. Compute the top-M left singular vectors, the top-M right singular vectors and the top-M singular value matrix and estimate initial values

$$\mathbf{W}^{init} = \mathbf{P}\boldsymbol{\Sigma}\mathbf{Q}^T, \mathbf{P} \in \mathbb{R}^{D \times M}, \boldsymbol{\Sigma} \in \mathbb{R}^{M \times M}, \mathbf{Q} \in \mathbb{R}^{T \times M}$$

$$\mathbf{U} = \mathbf{P}\boldsymbol{\Sigma}^{1/2} \;\&\; \boldsymbol{V} = \boldsymbol{\Sigma}^{1/2}\mathbf{Q}^T$$

3. Repeat until convergence

4. Update $\mathbf{U}$ with an alternating direction method of multipliers and an early stopping

$$\min_{\mathbf{U},\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3} \sum_{j=1}^{T} \frac{1}{N_j} L(\mathbf{y}_j, \mathbf{X}_j \mathbf{Z}_1 \mathbf{v}_j) + \gamma_1 \|\mathbf{Z}_2\|_1 + \gamma_2 \|\mathbf{Z}_3\|_{1,\infty}$$

$$s.t\ \mathbf{A}\mathbf{U} + \mathbf{B}\mathbf{Z} = \mathbf{0},$$

where $\mathbf{A} = \begin{bmatrix}\mathbf{I}_D \\ \mathbf{I}_D \\ \mathbf{I}_D\end{bmatrix}$, $\boldsymbol{B} = \boldsymbol{diag}(-\mathbf{I}_D, -\mathbf{I}_D, -\mathbf{I}_D)$, and $\mathbf{Z} = \begin{bmatrix}\mathbf{Z}_1 \\ \mathbf{Z}_2 \\ \mathbf{Z}_3\end{bmatrix}$

5. For $j = 1, \dots, T$, update $\mathbf{v}_j$ by solving a k-support norm regularized regression or logistic regression with an accelerated proximal gradient descent

$$\min_{\mathbf{v}} \frac{1}{N_j} L(\mathbf{y}_j, (\mathbf{X}_j \mathbf{U})\mathbf{v}) + \mu\left(\|\mathbf{v}\|_k^{sp}\right)^2$$

6. End Repeat

Performance bound

Reformulation

$$\min_{\mathbf{U} \in \mathcal{H}, \mathbf{v}_j \in \mathcal{F}} \frac{1}{NT} \sum_{j=1}^{T} L'(\mathbf{y}_j, \mathbf{X}_j \mathbf{U}\mathbf{v}_j)$$

where $\mathcal{H} = \{x \in \mathbb{R}^D \to (u_1^T x, \dots, u_M^T x) \in \mathbb{R}^M : \mathbf{u}_1, \dots, \mathbf{u}_M \in \mathbb{R}^D$, $\sum_{i=1}^{D} \|\mathbf{u}^i\|_1 \le \alpha_1, \sum_{i=1}^{D} \|\mathbf{u}^i\|_1 \le \alpha_2\}$, $\mathcal{F} = \left\{\mathbf{z} \in \mathbb{R}^M \to \mathbf{v}^T\mathbf{z} \in \mathbb{R} : \mathbf{v} \in \mathbb{R}^M, \left(\|\mathbf{v}\|_{sp}^k\right)^2 \le \beta^2\right\}$, and L' is the scaled loss function in [0,1]

Upper bound on the excess error from Maurer *et al.*, 2016

If $\alpha_1^2 \le M$, with probability at least $1-\delta$ the excess error is bounded by

$$\frac{1}{T}\sum_{j=1}^{T} \mathbb{E}\left[L'(\mathbf{y}_j, \mathbf{X}_j \hat{\mathbf{U}}\hat{\mathbf{v}}_j)\right] - \min_{\mathbf{U} \in \mathcal{H}, \mathbf{v}_j \in \mathcal{F}} \frac{1}{T}\sum_{j=1}^{T} \mathbb{E}\left[L'(\mathbf{y}_j, \mathbf{X}_j \mathbf{U}\mathbf{v}_j)\right]$$

$$\le c_1 \beta M \sqrt{\frac{\|\hat{C}(\bar{\mathbf{X}})\|_1}{NT}} + c_2 \beta \sqrt{\frac{\|\hat{C}(\bar{\mathbf{X}})\|_\infty}{N}} + \sqrt{\frac{8 \ln\left(\frac{2}{\delta}\right)}{NT}}$$

where $\hat{\mathbf{U}}, \hat{\mathbf{v}}_1, \dots \hat{\mathbf{v}}_T$ are the optimal solution of the reformulated problem, $\|\hat{C}(\bar{\mathbf{X}})\|_1 = \frac{1}{T}\sum_{j=1}^{T} tr\left(\hat{\Sigma}(\mathbf{X}_j)\right)$, $\|\hat{C}(\bar{\mathbf{X}})\|_\infty = \frac{1}{T}\sum_{j=1}^{T} \lambda_{max}\left(\hat{\Sigma}(\mathbf{X}_j)\right)$, and $\hat{\Sigma}(\mathbf{X}_j)$ is the empirical covariance of $\mathbf{X}_j$

Experiment - Benchmark

- LASSO
- L1+ Trace norm (Richard *et al.*, 2012)
- Multiplicative Multi-task Feature Learning (MMTFL) (Wang *et al.*, 2016)
- Clustered Multi-task Learning (CTML) (Zhou *et al.*, 2011)
- Group Overlap Multi-task Learning (GO-MTL) (Kumar and Daumé , 2012)

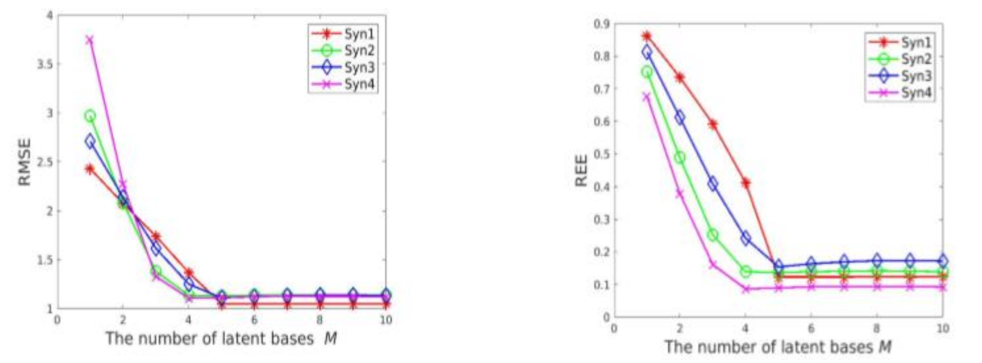
Experiment - Result

Evaluation measure

- Root mean squared error (RMSE), Relative estimation error (REE), Error rate (ER)

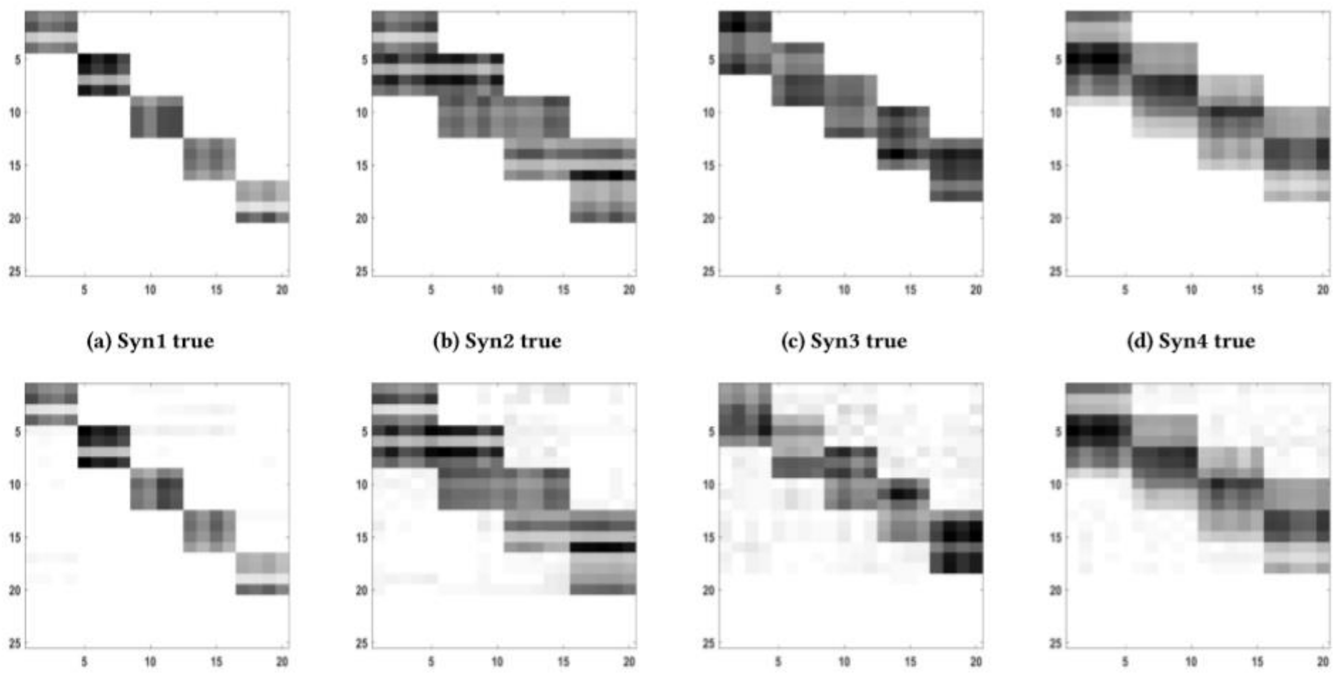
Synthetic Datasets

Synthetic	Measure	LASSO	L1+trace	MMTFL	CMTL	GO-MTL	VSTG-MTL $k=1$	VSTG-MTL $k=3$
Syn1	RMSE	1.4625	1.1585	1.1384	1.3170	1.0935	**1.0550**	1.0795
		± 0.1349	± 0.1585	± 0.0257	± 0.0298	± 0.0185	**± 0.0228**	± 0.0184
	REE	0.4155	0.2249	0.2089	0.3277	0.1737	0.1226	0.1536
		± 0.0595	± 0.0200	± 0.0169	± 0.0186	± 0.0165	± 0.0149	± 0.0128
Syn2	RMSE	1.6811	1.2639	1.2377	1.3720	1.1509	1.1067	1.1090
		± 0.1146	± 0.0418	± 0.0401	± 0.0497	± 0.0267	± 0.0282	± 0.0258
	REE	0.3703	0.2040	0.1921	0.2479	0.1488	0.1231	0.1230
		± 0.0441	± 0.0169	± 0.0152	± 0.0170	± 0.0122	± 0.0118	± 0.0095
Syn3	RMSE	1.5303	1.2244	1.1797	1.3470	1.1129	1.1013	**1.0068**
		± 0.0483	± 0.0320	± 0.0287	± 0.0334	± 0.0250	± 0.0244	**± 0.0201**
	REE	0.3801	0.2262	0.2001	0.2881	0.1565	0.1473	**0.1412**
		± 0.0328	± 0.0211	± 0.0168	± 0.0200	± 0.0148	± 0.0133	**± 0.0110**
Syn4	RMSE	1.7380	1.2673	1.2271	1.4418	1.1278	1.0863	**1.0618**
		± 0.1032	± 0.0312	± 0.0309	± 0.0402	± 0.0235	± 0.0225	**± 0.0211**
	REE	0.2729	0.1419	0.1302	0.1911	0.0945	0.0768	**0.0741**
		± 0.0365	± 0.0125	± 0.0111	0.0103	± 0.0087	± 0.0117	**± 0.0093**



(a) The number of latent bases M vs RMSE

(b) The number of latent bases M vs REE



(a) Syn1 true (b) Syn2 true (c) Syn3 true (d) Syn4 true

(e) Syn1 VSMTL k=1 (f) Syn2 VSMTL k=1 (g) Syn3 VSMTL k=3 (h) Syn4 VSMTL k=3

Real datasets

Dataset	Measure	LASSO	L1+Trace	MMTFL	CMTL	GO-MTL	VSTG-MTL k=opt
School exam	RMSE	12.0483	10.5041	10.1303	10.0170	10.1924	**9.8931**
		± 0.1738	± 0.1432	± 0.1291	± 0.1979	± 0.1331	**± 0.1103**
Parkinson		2.9177	1.0481	1.1079	1.0408	1.0231	1.0077
		± 0.0960	± 0.0243	± 0.0182	± 0.0229	± 0.0285	± 0.0191
Computer survey		2.3119	4.9493	1.7525	2.7562	1.9067	**1.6993**
		± 0.3997	± 2.1592	± 0.1237	± 0.6336	± 0.1864	**± 0.1053**
MNIST	ER	13.0200	17.9800	12.6000	12.3400	12.8400	**11.7000**
		± 0.7084	± 1.7574	± 0.8641	± 0.0199	1.2989	**± 1.4461**
USPS		12.8800	16.0200	11.3600	12.4400	12.9000	11.4800
		± 1.5061	± 1.2874	± 1.1462	± 0.0099	± 1.0842	± 1.0379

Conclusion

- Propose **VSTG-MTL** that performs both variable selection and task grouping based on low-rank factorization and sparsity
- Focus on possible correlation from representation learning
- Present a upper bound on the excess risk
- Source code : https://github.com/JunYongJeong/VSTG-MTL

Reference

- Kumar, A. and Daumé III, H. 2012. Learning Task Grouping and Overlap in Multi-Task Learning. ICML'12, 1383-1390
- Maurer, A., Pontil, M. and Romera-Paredes, B. 2016. The Benefit of Multitask Representation Learning. JMLR 17, 81 , 1–32
- Richard, E., Savalle, P.A. and Vayatis, N. 2012. Estimation of Simultaneously Sparse and Low Rank Matrices. ICML'12, 51–58
- Wang, X., Bi, J., Yu, S., Sun, J., and Song, M. 2016. Multiplicative Multitask Feature Learning. JMLR 17, 80 , 1–33
- Zhou, J., Chen, J., and Ye. J. 2011. Clustered Multi-Task Learning via Alternating Structure Optimization. NIPS'11. 702–710
- Argyriou, A., Foygel, R., and Srebro, N. 2012. Sparse Prediction with the k-Support Norm. NIPS'12,1457–1465.