Variable Selection and Task Grouping for Multi-Task Learning

Jun-Yong Jeong and Chi-Hyuck Jun
Industrial and Management Engineering, POSTECH

Introduction

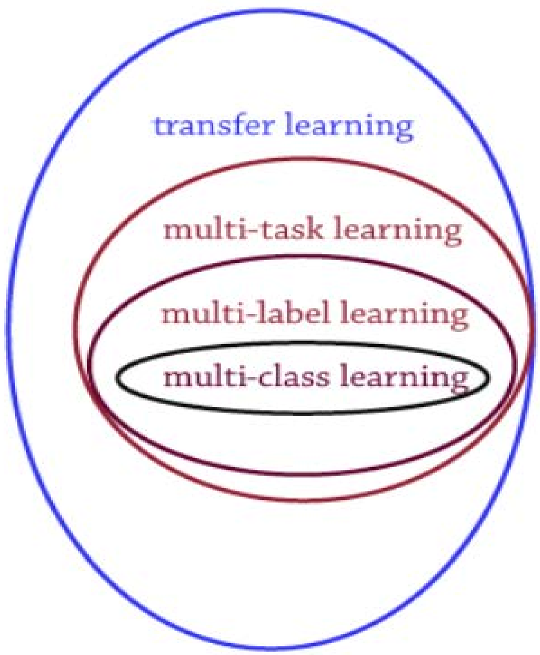
- **Multi-task learning**
  - Multiple related output variables (=Task)
  - Different observations for each output variable

Input variables			Output variables (=Tasks)		
X_1	…	X_D	Y_1	…	Y_T
2		40	10		?
3		23	20		11
4		100	?		15
1.5		10	?		9
2		53	17		?

Introduction

- **Multi-task learning**
  - Relationship to other problems

(figure from Zhou et al., 2012)

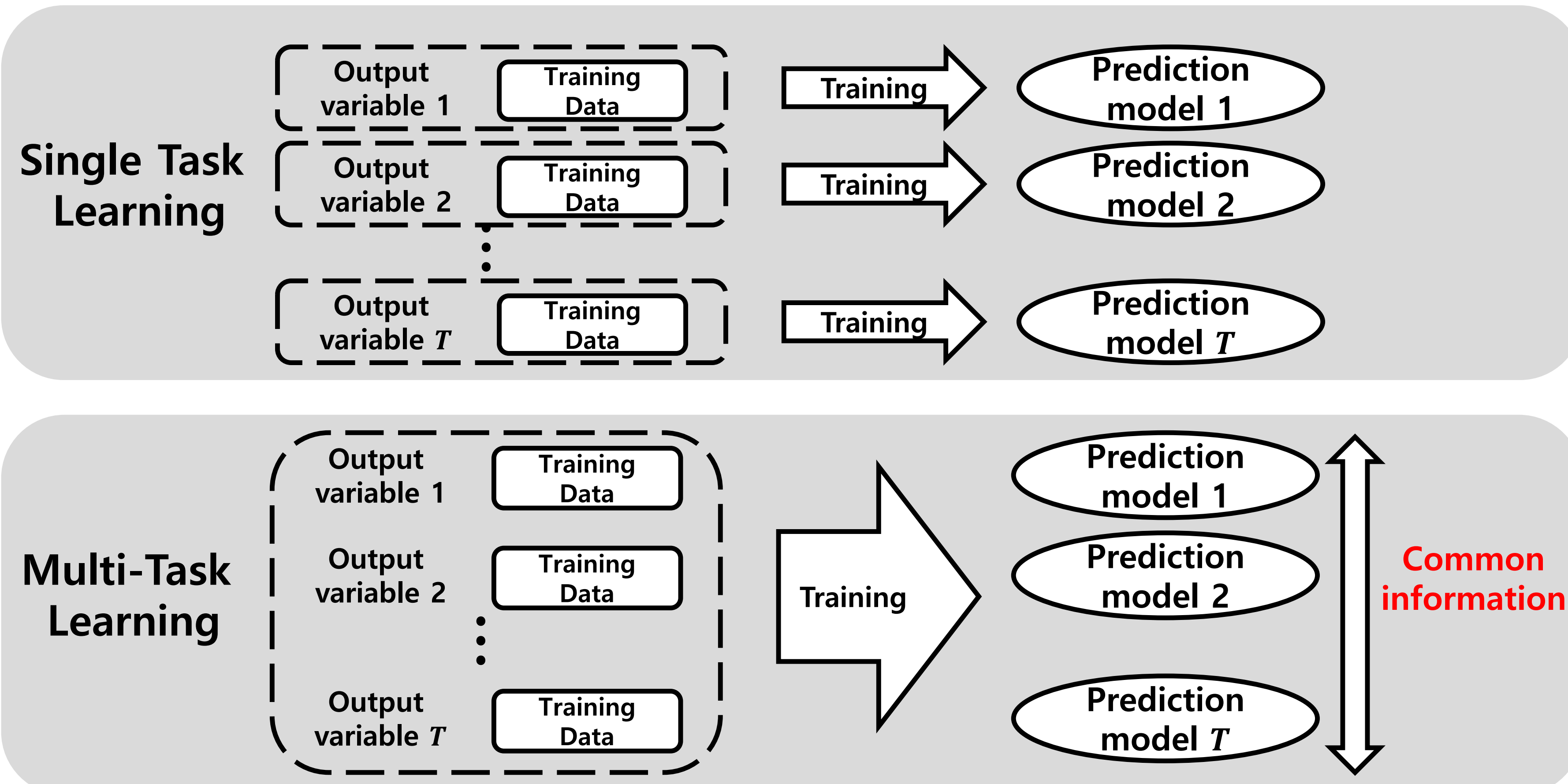


- Transfer Learning
  - Define source & target domains
  - Learn on the source domain
  - Generalize on the target domain
- Multi-task Learning
  - Model the task relatedness
  - Learn all tasks simultaneously
  - Tasks may have different data/features
- Multi-label Learning
  - Model the label relatedness
  - Learn all labels simultaneously
  - Labels share the same data/features
- Multi-class Learning
  - Learn the classes independently
  - All classes are exclusive

Introduction

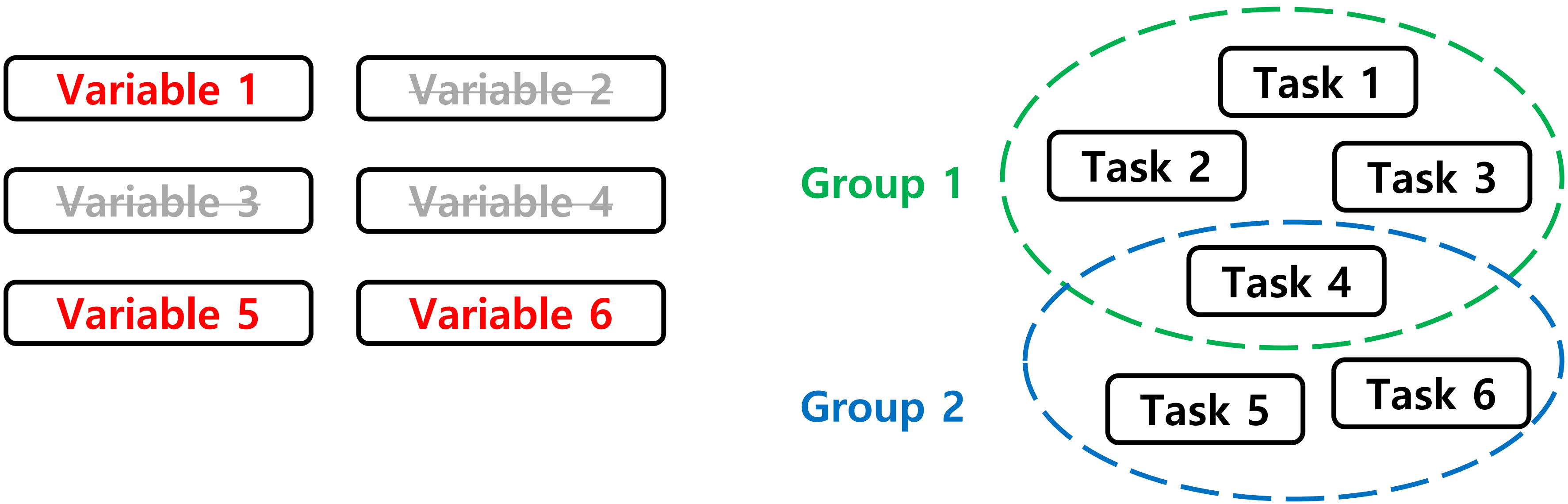
- **Multi-task learning**
  - Simlutaneous learning to share common information among prediction models

(figure from Zhou et al., 2012)



Proposed method

- **Problem and Purpose**
  - Multi-task regression/classification
  - T tasks (=output) and D input variables

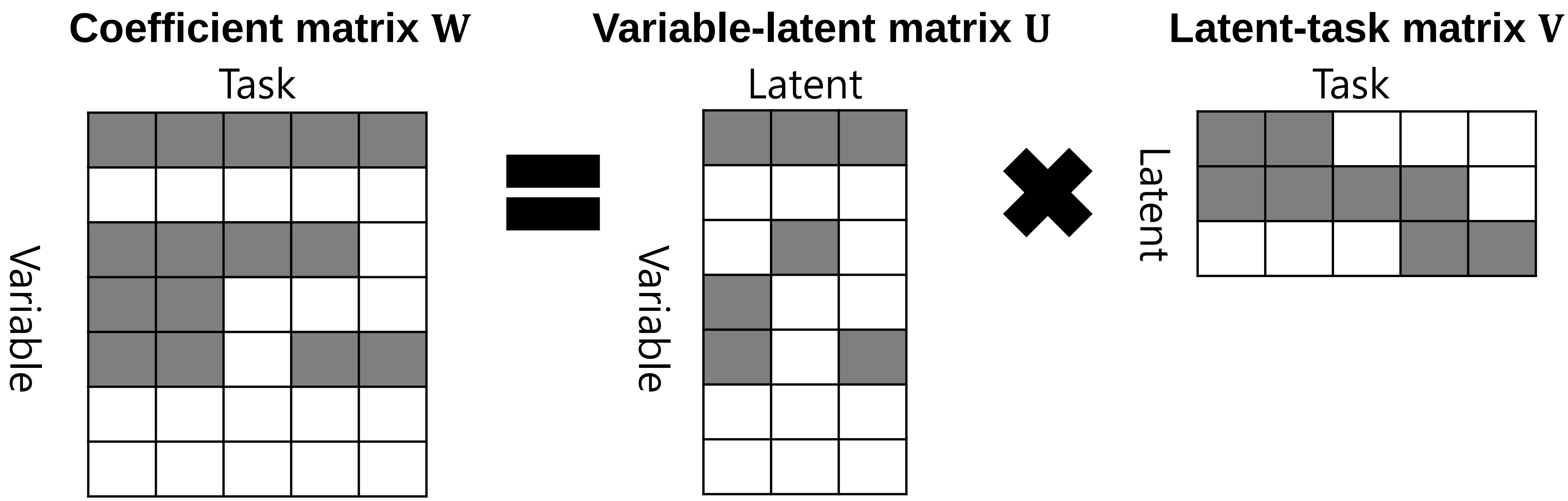


Proposed method

- **Main idea**
  - Linear model
  - Low-rank factorization & Sparsity

$$\mathbf{W} = \mathbf{UV} \in \mathbb{R}^{D\times T}$$

$$\mathbf{U} \in \mathbb{R}^{D\times M} \text{ and } \mathbf{V} \in \mathbb{R}^{M\times T}$$

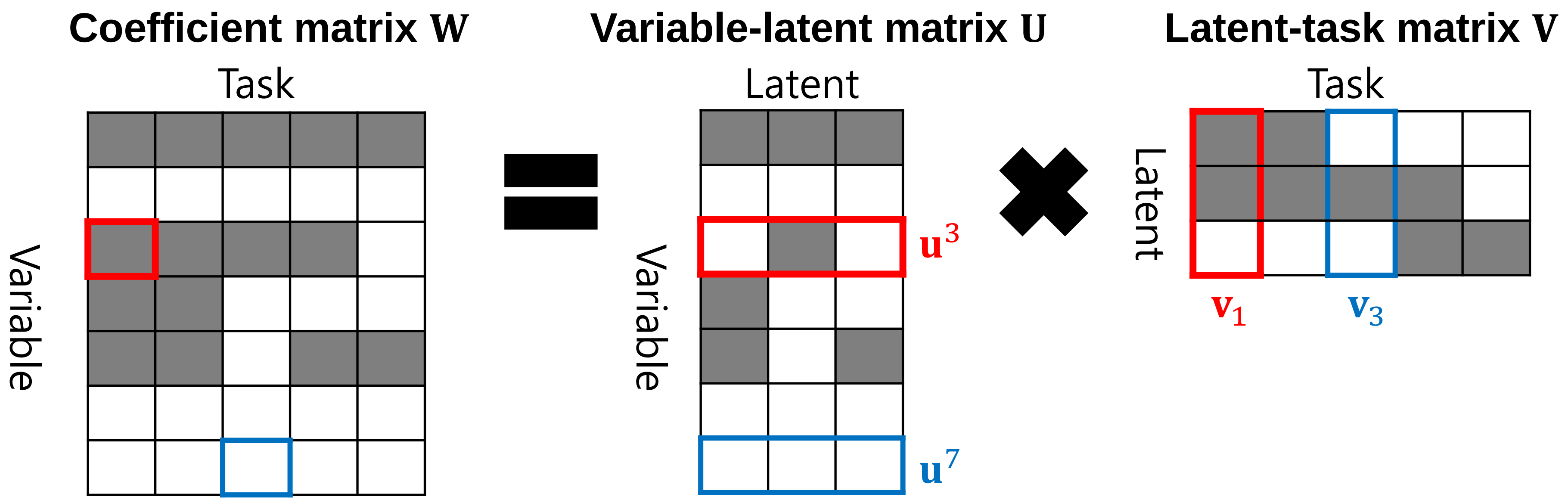


Proposed method

- **Variable selection**

  - Coefficient of ith variable for jth task $w_{ij} = \mathbf{u}^i \mathbf{v}_j \in \mathbb{R}$ $\left(\mathbf{u}^i \in \mathbb{R}^{1\times M}\ \&\ \mathbf{v}_j \in \mathbb{R}^M\right)$

  $\Rightarrow$ Impose sparsities between and within the rows of Variable-latent matrix $\mathbf{U}$ (Chen and Huang, 2012)

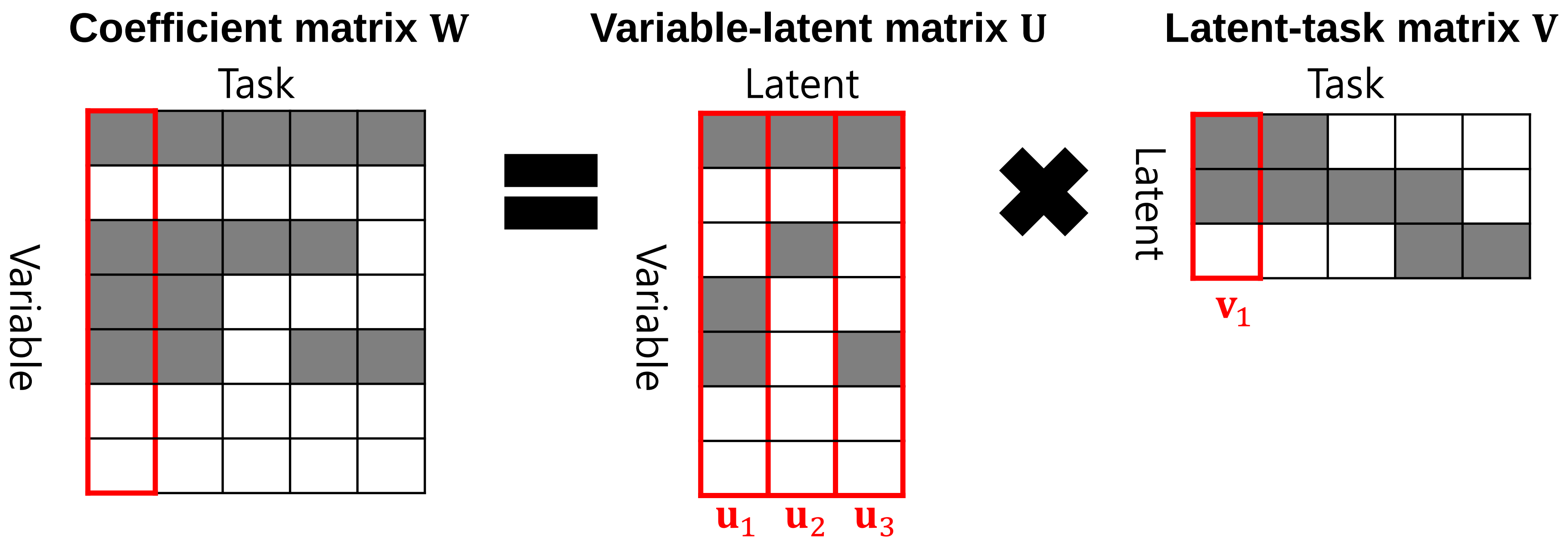


Proposed method

- **Task grouping**
  - Coefficient vector for jth task $\mathbf{w}_j = \mathbf{U}\mathbf{v}_j = \sum_{m=1}^{M} v_{mj}\mathbf{u}_m \in \mathbb{R}^D$

⇒ Task grouping by dependency on the basis vectors $\mathbf{u}_m$

⇒ Impose sparsity within the columns of Latent-task matrix $\mathbf{V}$ (Kumar and Daume, 2012)

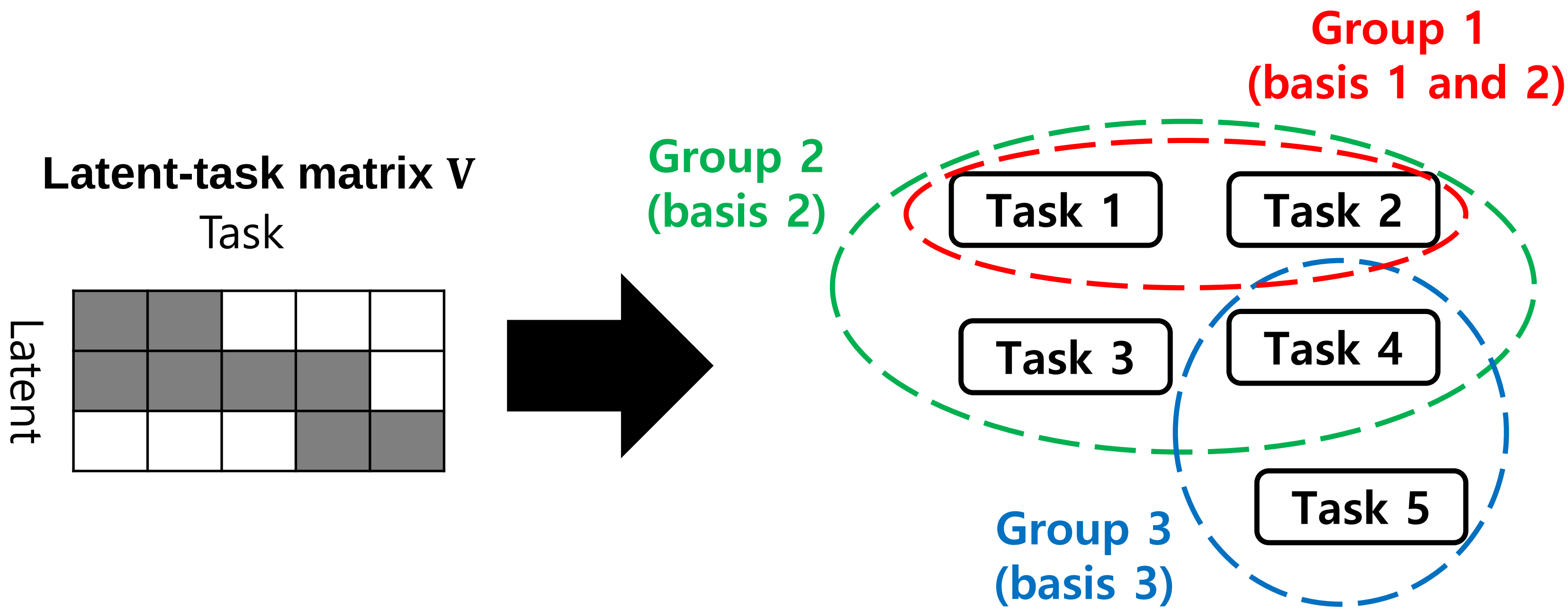


Proposed method

- **Task grouping**
  - Coefficient vector for jth task $\mathbf{w}_j = \mathbf{U}\mathbf{v}_j = \sum_{m=1}^{M} v_{mj}\mathbf{u}_m \in \mathbb{R}^D$

  $\Rightarrow$ Task grouping by dependency on the basis vectors $\mathbf{u}_m$

  $\Rightarrow$ Impose sparsity within the columns of Latent-task matrix $\mathbf{V}$ (Kumar and Daume, 2012)



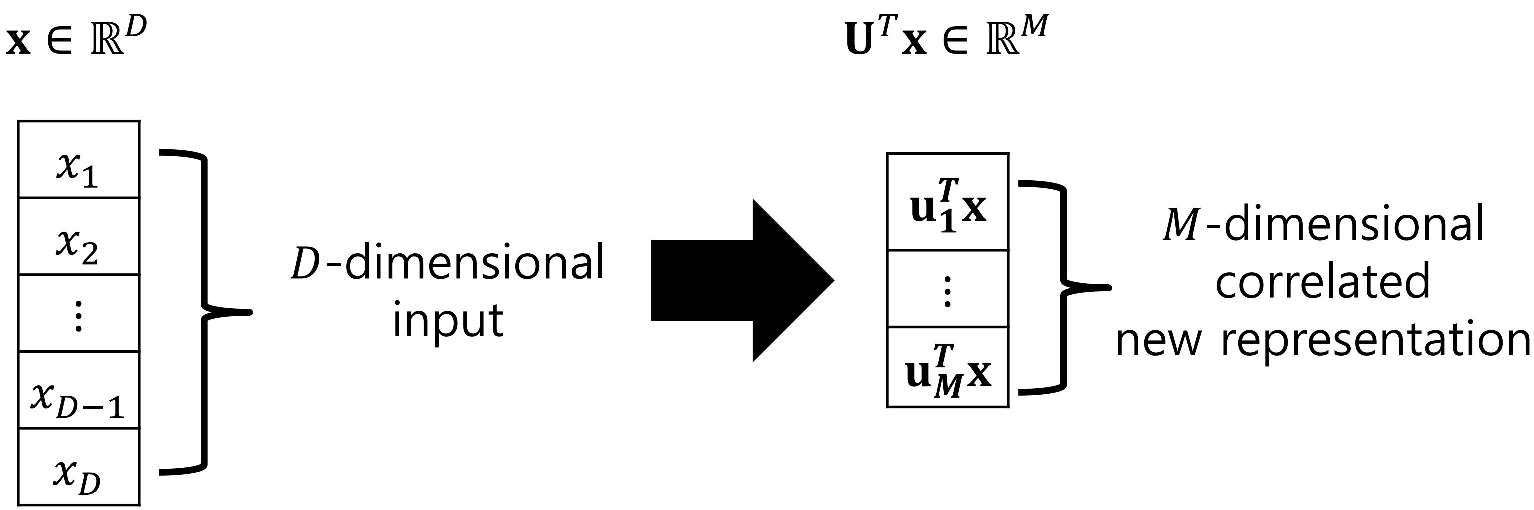
Proposed method

- **Task grouping**
  - Representation learning

$$\hat{y}_j(\mathbf{x}) = \left(\mathbf{w}_j\right)^T \mathbf{x} = \mathbf{v}_j^T \mathbf{U}^T \mathbf{x} = \mathbf{v}_j^T \left(\mathbf{U}^T \mathbf{x}\right)$$

$\mathbf{U}^T$**:** a linear transform from $\mathbb{R}^D$ to $\mathbb{R}^M$

$\mathbf{v}_j \in \mathbb{R}^M$: a coefficient vector in a new correlated representation



Proposed method

- **Optimization problem**

$$\min_{\mathbf{U},\mathbf{V}} \sum_{j=1}^{T} \frac{1}{N_j} L\left(y_j, \mathbf{X}_j \mathbf{U} \mathbf{v}_j\right)$$
$$s.t$$

C1: $\| \mathbf{U} \|_1 = \sum_{i=1}^{D} \| \mathbf{u}^i \|_1 \leq \alpha_1,$
C2: $\| \mathbf{U} \|_{1,\infty} = \sum_{i=1}^{D} \| \mathbf{u}^i \|_\infty \leq \alpha_2,$
C3: $\sum_{j=1}^{T} \left(\| \mathbf{v}_j \|_k^{sp}\right)^2 \leq \beta$

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

C1 & C2: L1,1 & L1,inf norm
⇒ impose sparsities between and within the row vector $\mathbf{u}^i$
⇒ perform variable selection

C3: Squared k-support norm (Argyriou et al., 2012)
⇒ impose sparsity within the column vector $\mathbf{v}_j$ while considering correlation
⇒ perform task grouping

Proposed method

Optimization procedure
Initialization based on single-task learning 1. Learn a ridge regression for each task to compute initial coefficient vector $$\mathbf{w}_j^{init} := \underset{\mathbf{w}}{\operatorname{argmin}} \frac{1}{N_j} L(\mathbf{y}_j, \mathbf{X}_j \mathbf{w}) + \sqrt{\gamma_1^2 + \gamma_2^2 + \mu^2} \parallel \mathbf{w} \parallel_2^2$$ $$\mathbf{W}^{\text{init}} := [\mathbf{w}_1^{init}, \dots, \mathbf{w}_T^{init}] \in \mathbb{R}^{D \times T}$$ 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 $$\mathbf{W}^{\boldsymbol{init}} = \mathbf{P\Sigma Q}^T, \mathbf{P} \in \mathbb{R}^{D \times M}, \mathbf{\Sigma} \in \mathbb{R}^{M \times M}, \mathbf{Q} \in \mathbb{R}^{T \times M}$$ $$\mathbf{U} = \mathbf{P\Sigma}^{\mathbf{1/2}} \ \& \ \mathbf{V} = \mathbf{\Sigma}^{\mathbf{1/2}} \mathbf{Q}^T$$ **## Alternating optimization** 3. Repeat until convergence 4. Update $\mathbf{U}$ with an ADMM and an early stopping $$\min_{\mathbf{U}, \mathbf{Z_1}, \mathbf{Z_2}, \mathbf{Z_3}} \sum_{j=1}^{T} \frac{1}{N_j} L(y_j, \mathbf{X}_j \mathbf{Z}_1 \mathbf{v}_j) + \gamma_1 \parallel \mathbf{Z}_2 \parallel_1 + \gamma_2 \parallel \mathbf{Z_3} \parallel_{1,\infty}$$ $$s.t \ \mathbf{AU} + \mathbf{BZ} = \mathbf{0},$$ where $\mathbf{A} = \begin{bmatrix} \mathbf{I}_D \\ \mathbf{I}_D \\ \mathbf{I}_D \end{bmatrix}, \mathbf{B} = \boldsymbol{diag}(-\mathbf{I}_D, -\mathbf{I}_D, -\mathbf{I}_D)$, 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 $$\min_{\mathbf{v}} \frac{1}{N_j} L(\mathbf{y}_j, (\mathbf{X}_j \mathbf{U}) \mathbf{v}) + \mu (\parallel \mathbf{v} \parallel_k^{sp})^2$$ 6. End Repeat

Proposed method

▪ Theoretical analysis

- 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

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

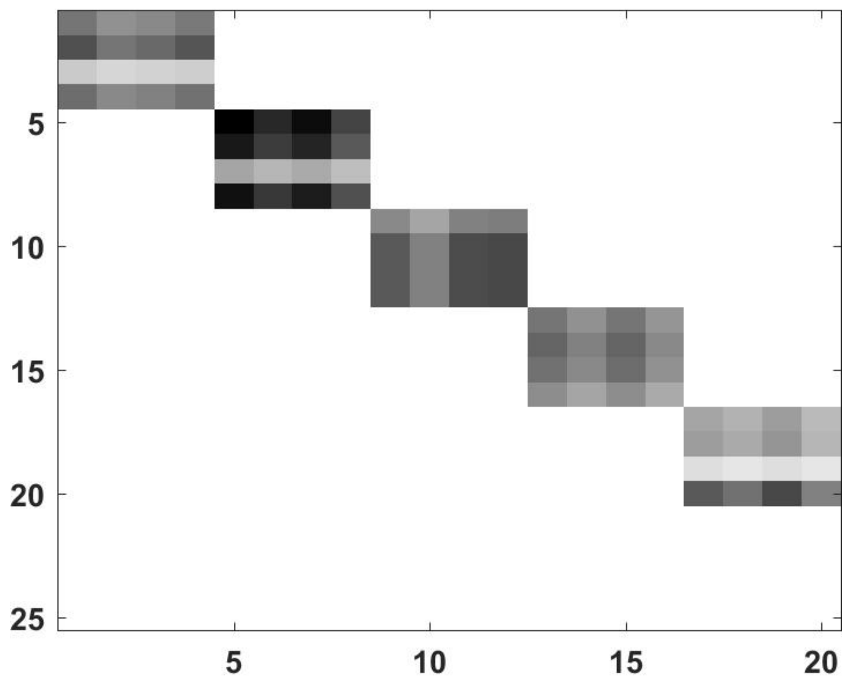
$$\le c_1\beta M\sqrt{\frac{\|\hat{C}(\overline{\mathbf{X}})\|_1}{NT}}+c_2\beta\sqrt{\frac{\|\hat{C}(\overline{\mathbf{X}})\|_\infty}{N}}+\sqrt{\frac{8\ln\left(\frac{2}{\delta}\right)}{NT}}$$

Where L' is the scaled loss function, $\widehat{\mathbf{U}},\hat{\mathbf{v}}_1,\dots\hat{\mathbf{v}}_T$ are the optimal solution, $\|\hat{C}(\overline{\mathbf{X}})\|_1=\frac{1}{T}\sum_{j=1}^{T}tr\left(\hat{\Sigma}\left(\mathbf{X}_j\right)\right)$, $\|\hat{C}(\overline{\mathbf{X}})\|_\infty=\frac{1}{T}\sum_{j=1}^{T}\lambda_{max}\left(\hat{\Sigma}\left(\mathbf{X}_j\right)\right)$, and $\hat{\Sigma}\left(\mathbf{X}_j\right)$ is the empirical covariance of $\mathbf{X}_j$

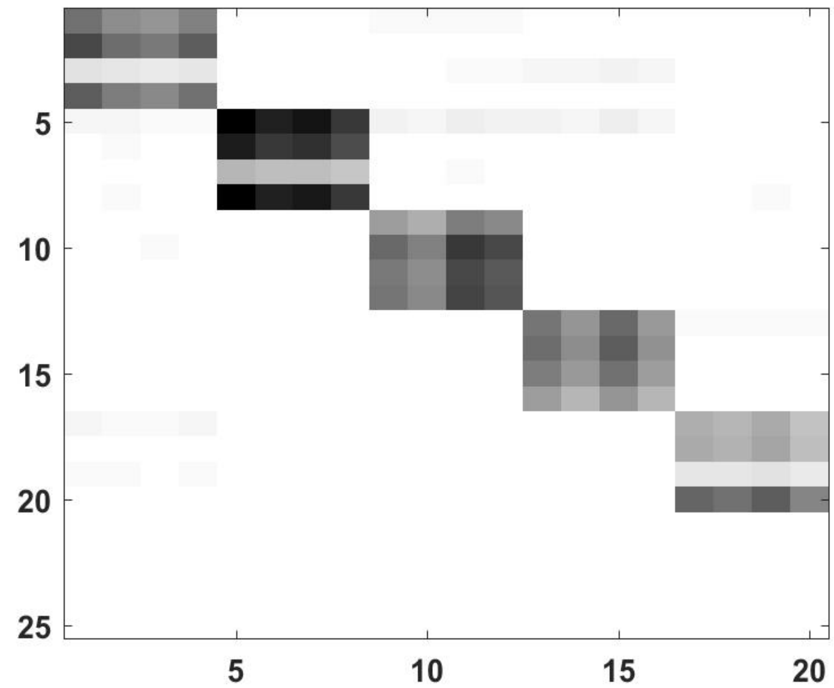
Experiment

- **Simulation study**
  - True model: $\mathbf{W}^* = \mathbf{U}^*\mathbf{V}^* \in \mathbb{R}^{25\times 20}$ & $y_j = \mathbf{x}^T\mathbf{w}_j^* + N(0,1)$
  - Case 1. No correlation & disjoint group

True coefficient matrix $\mathbf{W}^*$



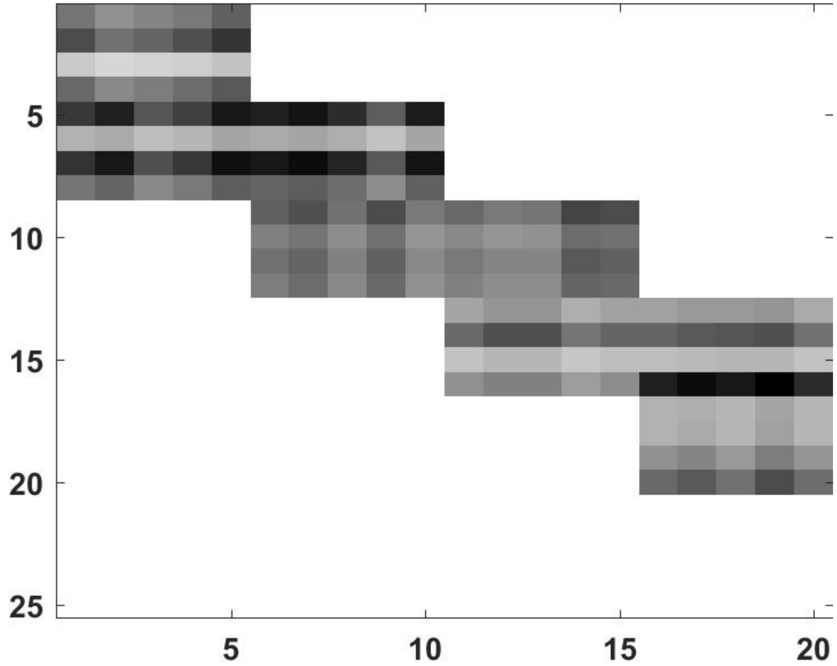
Estimated coefficient matrix $\widehat{\mathbf{W}}$



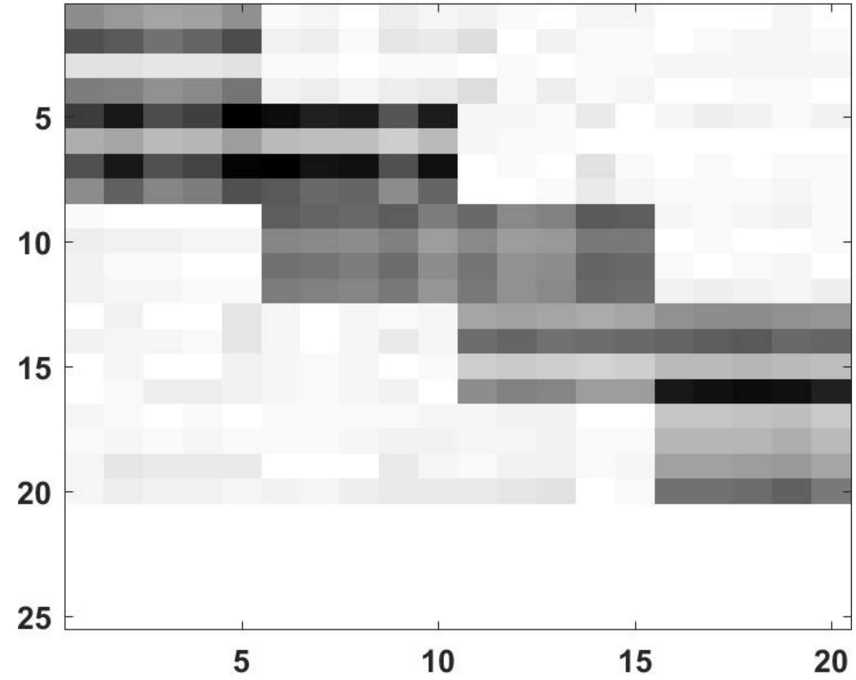
Experiment

- **Simulation study**
  - True model: $\mathbf{W}^* = \mathbf{U}^*\mathbf{V}^* \in \mathbb{R}^{25\times 20}$ & $y_j = \mathbf{x}^T\mathbf{w}_j^* + N(0,1)$
  - Case 2. No correlation & overlapping group

True coefficient matrix $\mathbf{W}^*$



Estimated coefficient matrix $\widehat{\mathbf{W}}$



Experiment

- **Benchmark datasets**

Datasets		# of input variables (D)	# of tasks (T)	# of total observations	Train/Test
Regression	School exam (Goldstein, 1991)	26	139	15,362	75%/25%
	Parkinson (Tsanas et al., 2010)	19	42	5875	75%/25%
	Computer survey (Lenk et al., 1996)	13	190	20	75%/25%
Classification	MNIST (Lecun et al., 1998)	28×28 $\Rightarrow$ 64 by PCA	10	70,000	1000/500
	USPS (Hull, 1994)	16×16 $\Rightarrow$ 87 by PCA	10	9,298	1000/500

Experiment

- **Benchmark datasets – Regression**
  - Root mean squared error

Method		**School exam**	**Parkinson**	**Computer survey**
Single-task linear	LASSO	12.0483 (0.1738)	2.9177 (0.0960)	2.3199 (0.3997)
Multi-task linear	L1+TRACE (Richard et al., 2012)	10.5041 (0.1432)	1.0481 (0.0243)	4.9493 (2.1592)
	MMTFL (Wang et al., 2016)	10.1303 (0.1291)	1.1079 (0.0182)	1.7525 (0.1237)
	CTML (Zhou et al., 2011)	10.0170 (0.1979)	1.0408 (0.0229)	2.7562 (0.6336)
	GO-MTL (Kumar and Daume, 2012)	10.1924 (0.01331)	1.0231 (0.0285)	1.9067 (0.1864)
	Proposed (Jeong and Jun, 2018)	**9.8931 (0.1103)**	1.0077 (0.0191)	**1.6993 (0.1053)**

Experiment

- **Benchmark datasets – Classification**
  - Accuracy

Method		MNIST	USPS
Single-task linear	LASSO	13.0200 (0.7084)	12.8800 (1.5061)
Multi-task linear	L1+TRACE (Richard et al., 2012)	17.9800 (1.7574)	16.0200 (1.2874)
	MMTFL (Wang et al., 2016)	12.6000 (0.8641)	11.3600 (1.1462)
	CTML (Zhou et al., 2011)	12.3400 (0.0199)	12.4400 (0.0099)
	GO-MTL (Kumar and Daume, 2012)	12.8400 (1.2989)	12.9000 (1.0842)
	Proposed (Jeong and Jun, 2018)	**11.7000 (1.4461)**	11.4800 (1.0379)

Conclusion

- **Summary**
  - Linear model for multi-task regression and classification
  - Variable selection and Task grouping
  - Lower bound on the excess error

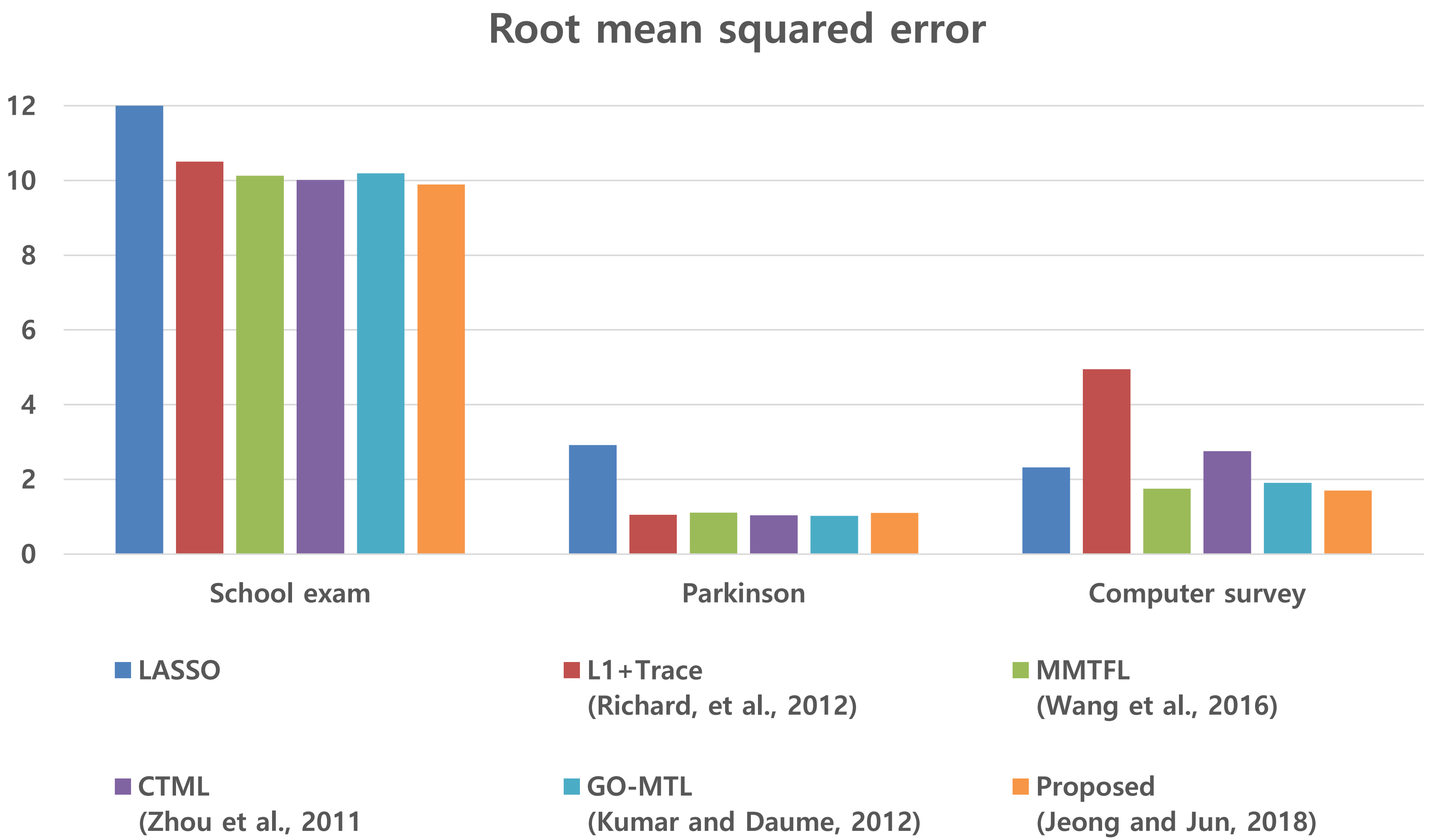
- **Future work**
  - Slow convergence rate of ADMM $O(1/\epsilon)$

  ⇒ Apply a proximal alternating linearized minimization (Bolte et al., 2014)

  & Compute a proximal operator of $\ell_1 + \ell_\infty$ norm

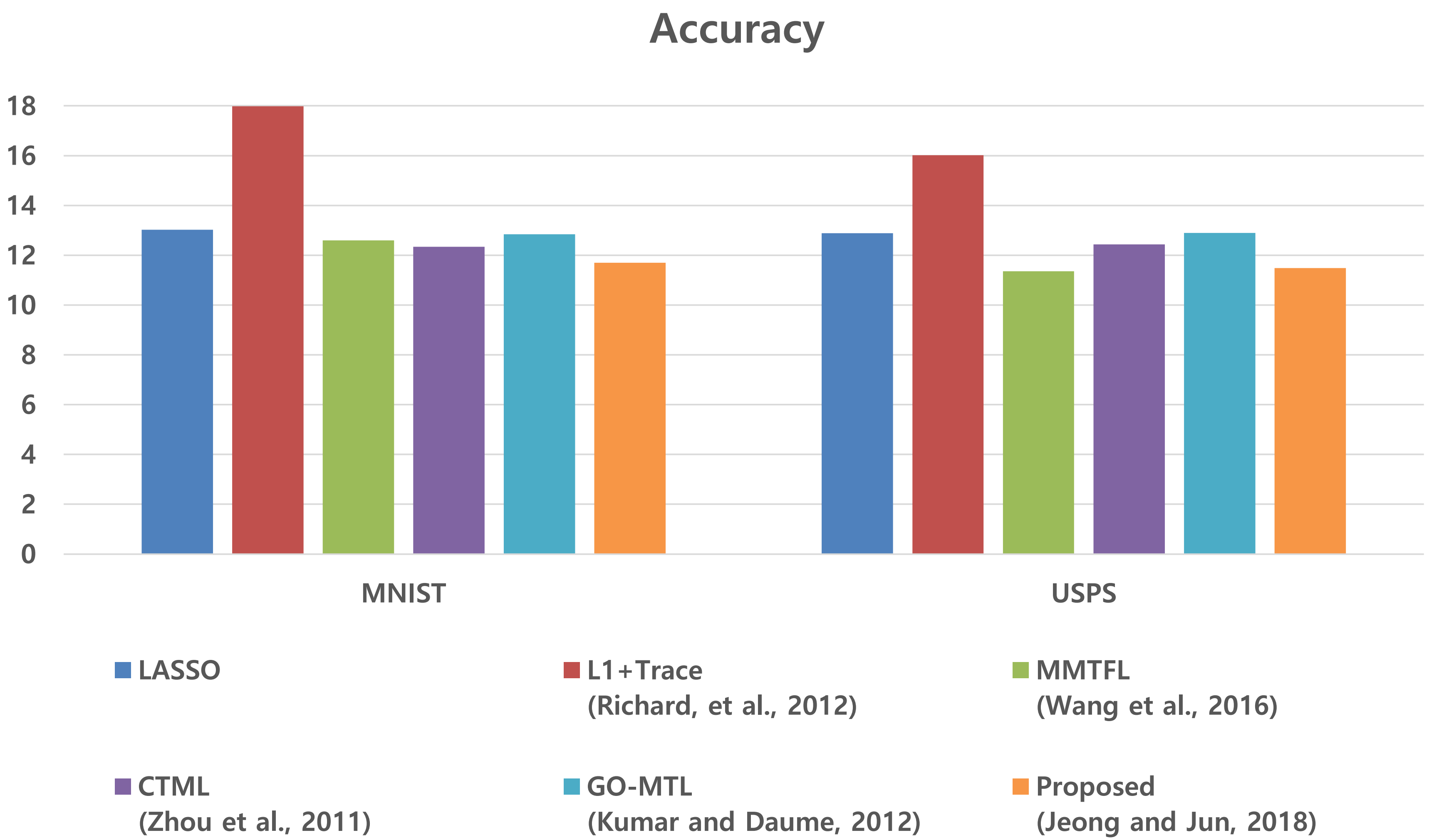
Experiment

- **Benchmark datasets – Regression**



Experiment

- **Benchmark datasets – Classification**



Reference

- Andreas Argyriou, Rina Foygel, and Nathan Srebro. 2012. Sparse Prediction with the k-Support Norm. In Advances in Neural Information Processing Systems 25 (NIPS'12), F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.).Curran Associates, Inc., 1457–1465.
- Jerome Bolte, Shoham Sabach, and Marc Teboulle. 2014. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. Mathematical Programming 146, 1 (01 Aug 2014), 459–494.
- Lisha Chen and Jianhua Z.Huang. 2012. Sparse Reduced-Rank Regression for Simultaneous Dimension Reduction and Variable Selection. J. Amer. Statist. Assoc. 107, 500 (2012), 1533–1545.
- Harvey Goldstein. 1991. Multilevel Modelling of Survey Data. Journal of the Royal Statistical Society. Series D (The Statistician) 40, 2 (1991), 235–244.
- A database for handwritten text recognition research. IEEE Transactions on Pattern Analysis and Machine Intelligence 16, 5 (May 1994), 550–554.
- Abhishek Kumar and Hal Daume III. 2012. Learning Task Grouping and Overlap in Multi-Task Learning. In Proceedings of the 29th International Conference on Machine Learning (ICML'12), John Langford and Joelle Pineau (Eds.). Omnipress, NY, USA, 1383–1390.
- Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient based learning applied to document recognition. Proc. IEEE 86, 11 (Nov 1998), 2278–2324.

Reference

- Peter J. Lenk, Wayne S. DeSarbo, Paul E. Green, and Martin R. Young. 1996. Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs. Marketing Science 15, 2 (1996), 173–191.
- Emile Richard, Pierre-Andre Savalle, and Nicolas Vayatis. 2012. Estimation of Simultaneously Sparse and Low Rank Matrices. In Proceedings of the 29th International Conference on Machine Learning (ICML'12). Omnipress, USA, 51–58
- Xin Wang, Jinbo Bi, Shipeng Yu, Jiangwen Sun, and Minghu Song. 2016. Multiplicative Multitask Feature Learning. Journal of Machine Learning Research 17, 80 (2016), 1–33.
- Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. 2016. The Benefit of Multitask Representation Learning. Journal of Machine Learning Research 17, 81 (2016), 1–32
- Athanasios Tsanas, Max A. Little, Patrick E. McSharry, and Lorraine O. Ramig. 2010. Accurate Telemonitoring of Parkinson's Disease Progression by Noninvasive Speech Tests. IEEE Transactions on Biomedical Engineering 57, 4 (April 2010), 884–893.
- Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2011. Clustered Multi-Task Learning via Alternating Structure Optimization. In Advances in Neural Information Processing Systems 24 (NIPS'11). Curran Associates, Inc., 702–710.
- Zhou, J., Chen, J., & Ye, J. (2012). Multi-Task Learning Theory: Theory, Algorithms, and Applications. In *ICDM 2012 tutorial*.