Efficient Multitask Feature and Relationship Learning

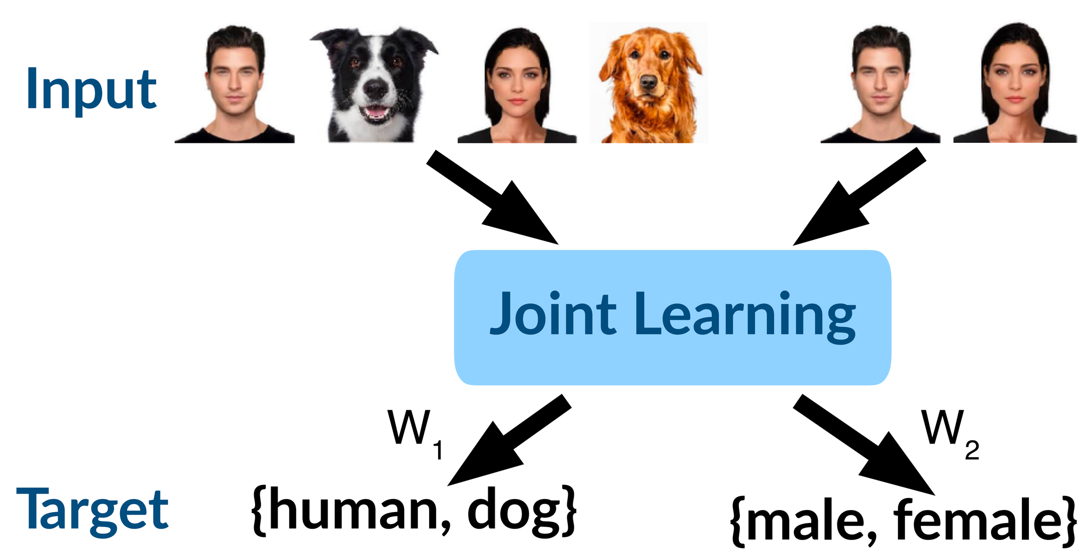
Han Zhao[◇], Otilia Stretcu[◇] Alex Smola[†] Geoff Gordon[‡]

{han.zhao, ostretcu}@cs.cmu.edu [†]alex@smola.org [‡]geoff.gordon@microsoft.com

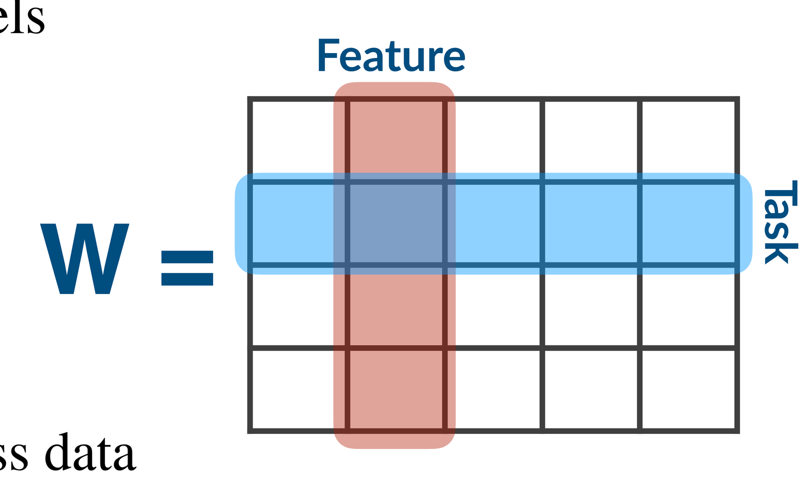
[◇]Carnegie Mellon University [†]Amazon [‡]Microsoft Research

Motivation

Multitask Learning:



- Multiple linear regression models
- Weight matrix W:
  - rows = tasks
  - columns = features
- Goal:
  - Joint learning multiple tasks
  - Better generalization with less data
  - Find correlation between tasks/features



Formulation

Empirical Bayes with prior:

$$W \mid \xi, \Omega_1, \Omega_2 \sim \left(\prod_{i=1}^{m} \mathcal{N}(\mathbf{w}_i \mid \mathbf{0}, \xi_i \mathbf{I}_d)\right) \cdot \mathcal{MN}_{d\times m}(W \mid \mathbf{0}_{d\times m}, \Omega_1, \Omega_2)$$

- $\mathcal{MN}_{d\times m}(W \mid \mathbf{0}_{d\times m}, \Omega_1, \Omega_2)$ is matrix-variate normal distribution
- $\Omega_1 \in \mathbb{S}^d_{++}$, covariance matrix over features
- $\Omega_2 \in \mathbb{S}^m_{++}$, covariance matrix over tasks
- $W \in \mathbb{R}^{d\times m}$, weight matrix

Maximum marginal-likelihood with empirical estimators:

$$\begin{aligned} \underset{W,\Sigma_1,\Sigma_2}{\text{minimize}} \quad & ||Y - XW||_F^2 + \eta||W||_F^2 + \rho||\Sigma_1^{1/2} W \Sigma_2^{1/2}||_F^2 \\ & - \rho(m \log|\Sigma_1| + d \log|\Sigma_2|) \\ \text{subject to} \quad & lI_d \preceq \Sigma_1 \preceq uI_d, lI_m \preceq \Sigma_2 \preceq uI_m \end{aligned}$$

where $\Sigma_1 := \Omega_1^{-1}$, $\Sigma_2 := \Omega_2^{-1}$.

- Multi-convex in W, Σ_1, Σ_2

Nonlinear extension:

- Replace the feature matrix X with the output of a neural network $g(\mathbf{x}; \theta)$ with learnable parameters θ.
- Estimate W and θ using backpropagation.
- Optimize the two covariance matrices using our proposed approach.

Optimization Algorithm

Solvers for W when Σ_1, Σ_2 are fixed:

$$\underset{W}{\text{minimize}} \quad h(W) \triangleq ||Y - XW||_F^2 + \eta||W||_F^2 + \rho||\Sigma_1^{1/2} W \Sigma_2^{1/2}||_F^2$$

Three different solvers:

- A closed form solution with $O(m^3 d^3 + mnd^2)$ complexity:

$$\text{vec}(W^*) = \left(I_m \otimes (X^T X) + \eta I_{md} + \rho \Sigma_2 \otimes \Sigma_1\right)^{-1} \text{vec}(X^T Y)$$

- Gradient computation:

$$\nabla_W h(W) = X^T(Y - XW) + \eta W + \rho \Sigma_1 W \Sigma_2$$

  Conjugate gradient descent with $O(\sqrt{\kappa}\log(1/\varepsilon)(m^2 d + md^2))$ complexity, κ is the condition number, ε is the approximation accuracy.

- Sylvester equation $AX + XB = C$ using the Bartels-Stewart solver. The first-order optimality condition:

$$\Sigma_1^{-1}(X^T X + \eta I_d)W + W(\rho\Sigma_2) = \Sigma_1^{-1} X^T Y$$

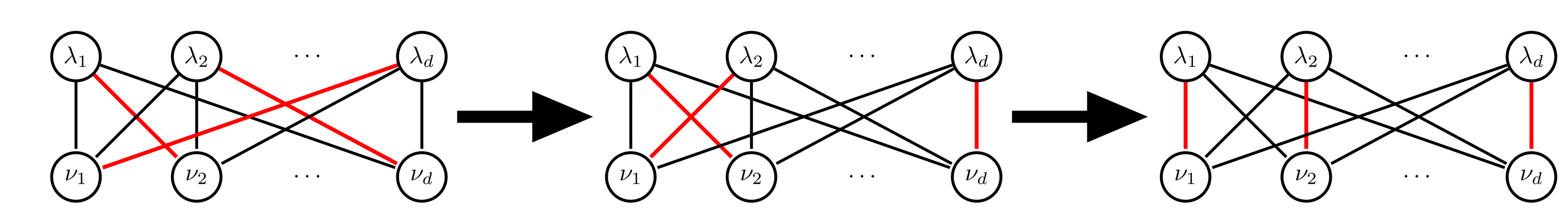
  Exact solution for W computable in $O(m^3 + d^3 + nd^2)$ time.

Solvers for Σ_1 and Σ_2 when W is fixed:

$$\underset{\Sigma_1}{\text{minimize}} \quad \text{tr}(\Sigma_1 W \Sigma_2 W^T) - m \log|\Sigma_1|, \qquad \text{subject to} \quad lI_d \preceq \Sigma_1 \preceq uI_d$$

$$\underset{\Sigma_2}{\text{minimize}} \quad \text{tr}(\Sigma_1 W \Sigma_2 W^T) - d \log|\Sigma_2|, \qquad \text{subject to} \quad lI_d \preceq \Sigma_2 \preceq uI_d$$

Exact solution by reduction to minimum-weight perfect matching:



Algorithms:

Input: W, Σ_2 and l, u.

1: $[V, \nu] \leftarrow \text{SVD}(W \Sigma_2 W^T)$.
2: $\lambda \leftarrow \mathbb{T}_{[l,u]}(m/\nu)$.
3: $\Sigma_1 \leftarrow V \text{diag}(\lambda) V^T$.

Input: W, Σ_1 and l, u.

1: $[V, \nu] \leftarrow \text{SVD}(W^T \Sigma_1 W)$.
2: $\lambda \leftarrow \mathbb{T}_{[l,u]}(d/\nu)$.
3: $\Sigma_2 \leftarrow V \text{diag}(\lambda) V^T$.

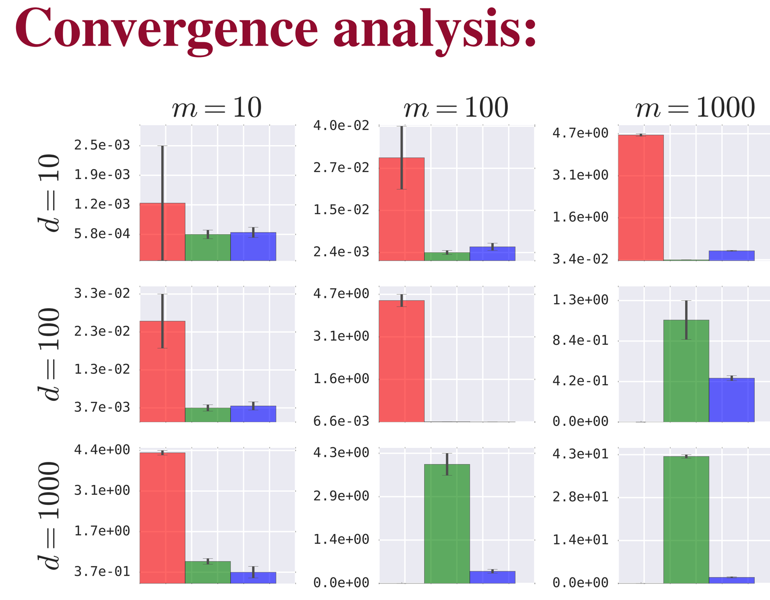
- Exact solution only requires one SVD
- Time complexity: $O(\max\{dm^2, md^2\})$

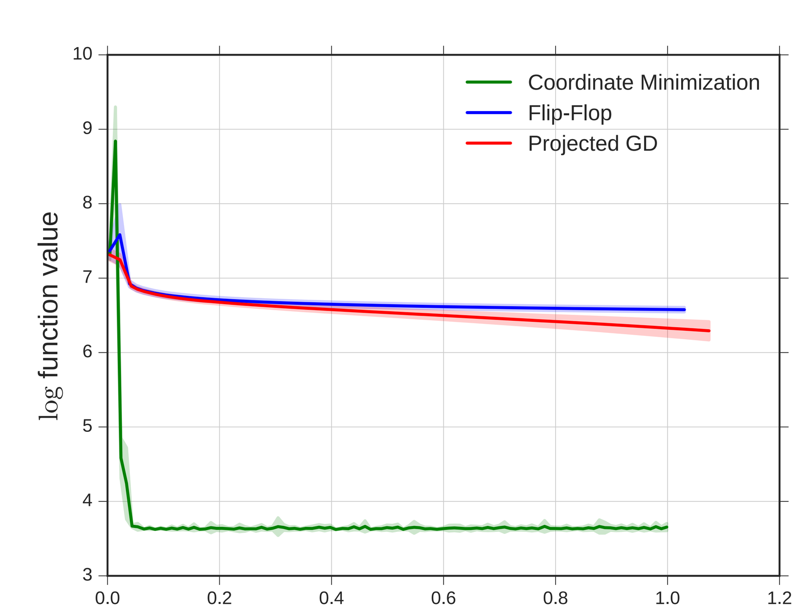
Experiments

Datasets:

- Synthetic data:
  - Randomly sample 10^4 instances, shared among all the tasks.
  - Gradually increase the dimension of features, d, and the number of tasks, m, to test scalability.
- Robot data (SARCOS):
  - $d = 21$ (7 joint positions, 7 joint velocities, 7 joint accelerations), $m = 7$ (7 joint torques).
  - 44,484 train instances, 4,449 test instances.
- School data:
  - $d = 27$, $m = 139$, $n = 15,362$ instances.
  - Goal: predict student scores.

Convergence analysis:



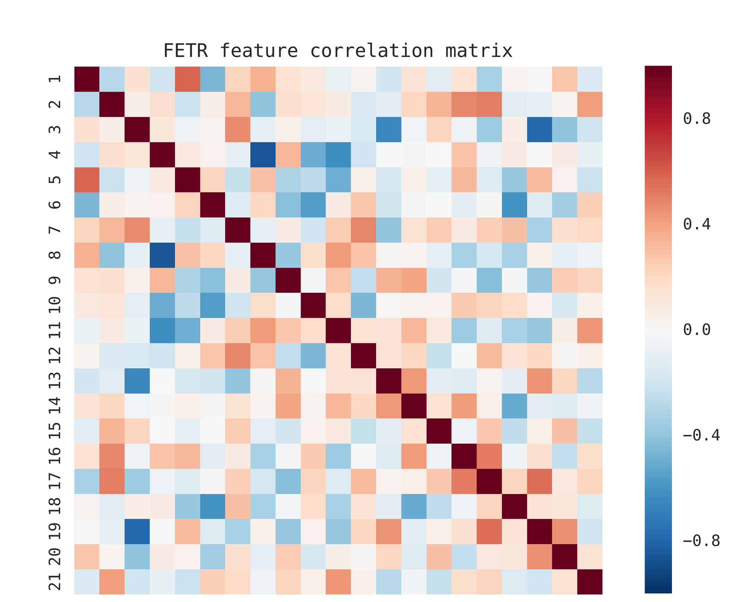


- The closed form solution does not scale when $md \geq 10^4$.

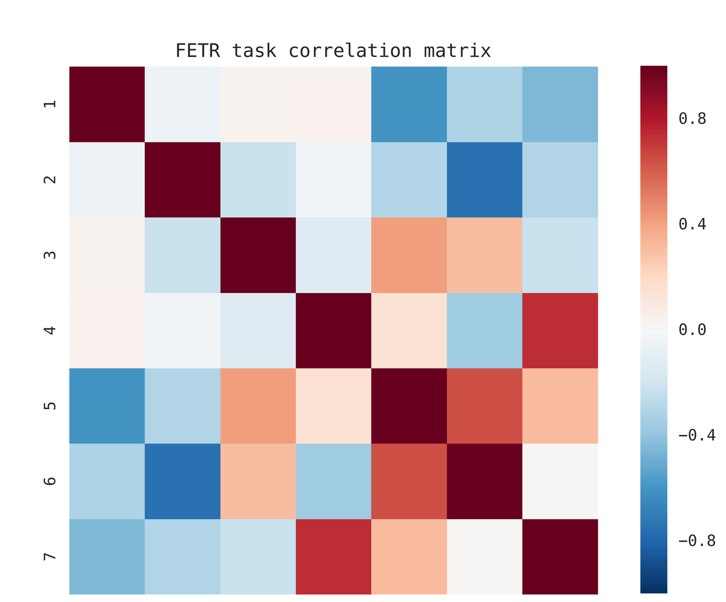
Results (mean squared error):

Method	SARCOS							School
	1st	2nd	3rd	4th	5th	6th	7th	
STL	31.40	22.90	9.13	10.30	0.14	0.84	0.46	0.9882 ± 0.0196
MTFL	31.41	22.91	9.13	10.33	0.14	**0.83**	0.45	0.8891 ± 0.0380
MTRL	31.09	22.69	**9.08**	9.74	0.14	**0.83**	0.44	0.9007 ± 0.0407
MTFRL	31.13	**22.60**	9.10	9.74	**0.13**	**0.83**	0.45	0.8451 ± 0.0197
FETR	**31.08**	22.68	**9.08**	**9.73**	**0.13**	**0.83**	**0.43**	$\mathbf{0.8134 \pm 0.0253}$
STL-NN	24.81	17.20	8.97	8.36	0.13	0.72	0.34	–
MT-NN	12.01	10.54	5.02	7.15	0.09	0.70	0.27	–
MTFRL-NN	11.02	9.51	4.99	7.11	**0.08**	0.62	0.27	–
FETR-NN	**10.77**	**9.34**	**4.95**	**7.01**	**0.08**	**0.59**	**0.24**	–

Feature covariance matrix and task covariance matrix:



(a) Covariance matrix over features.



(b) Covariance matrix over tasks.