Empirical Bayes with prior:
\[ W | \xi, \Omega_1, \Omega_2 \sim \mathcal{M}(\mathbf{w}_i | \mathbf{0}, \Sigma_{d \times m}) \cdot \mathcal{M}(\mathbf{w}_i | \mathbf{0}_{d \times m}, \Omega_1, \Omega_2) \]

Maximum marginal-likelihood with empirical estimators:
\[
\min_{w_{d \times m}} \|Y - XW\|^2_F + \eta \|W\|^2_F + \rho \|\Sigma_1^{1/2}W \Sigma_2^{1/2}\|^2_F.
\]
subject to
\[ H_d \leq \Sigma_1 \leq uH_d, H_m \leq \Sigma_2 \leq uH_m \]
where
\[ \Sigma_1 := \Omega_1^{-1}, \Sigma_2 := \Omega_2^{-1} \]

- Multi-convex in \( W, \Sigma_1, \Sigma_2 \)

Nonlinear extension:
- Replace the feature matrix \( X \) with the output of a neural network \( g(\xi; \theta) \) with learnable parameters \( \theta \).
- Estimate \( W \) and \( \theta \) using backpropagation.
- Optimize the two covariance matrices using our proposed approach.

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 \( md \geq 10^4 \).

Optimization Algorithm

### Solvers for \( W \) when \( \Sigma_1, \Sigma_2 \) are fixed:
\[
\minimize \ h(W) \triangleq \|Y - XW\|^2_F + \eta \|W\|^2_F + \rho \|\Sigma_1^{1/2}W \Sigma_2^{1/2}\|^2_F.
\]

Three different solvers:
- A closed form solution with \( O(md^2 + mnd^2) \) complexity:
  \[
  \text{vec}(W^*) = (I_m \otimes (X^T X) + \eta I_m + \rho \Sigma_1 \otimes \Sigma_2)^{-1} \text{vec}(X^T Y).
  \]
- Gradient computation:
  \[
  \nabla_W h(W) = X^T (Y - XW) + \eta W + \rho \Sigma_1 W \Sigma_2.
  \]
Conjugate gradient descent with \( O(\sqrt{k} \log(1/\varepsilon)(md^2 + mnd^2)) \) complexity, \( \varepsilon \) is the condition number, \( m \) 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 \( O(m^3 + d^3 + md^2) \) time.

### Solvers for \( \Sigma_1 \) and \( \Sigma_2 \) when \( W \) is fixed:
\[
\minimize \ tr(\Sigma_1 W \Sigma_2 W^T) - m \log |\Sigma_1|,\quad \text{subject to}\quad H_d \leq \Sigma_1 \leq uH_d.
\]
\[
\minimize \ tr(\Sigma_1 W \Sigma_2 W^T) - d \log |\Sigma_2|, \quad \text{subject to}\quad H_d \leq \Sigma_2 \leq uH_d.
\]

Exact solution by reduction to minimum-weight perfect matching:

Algorithms:
- **Input:** \( W, \Sigma_1 \) and \( l, u \).
  1. \([V, \nu] \leftarrow \text{SVD}(W \Sigma_1 W^T)\).
  2. \( \lambda \leftarrow \max_{d, \nu}(d/\nu)\).
  3. \( \Sigma_1 \leftarrow V \text{diag}(\lambda) V^T\).
- **Input:** \( W, \Sigma_2 \) and \( l, u \).
  1. \([V, \nu] \leftarrow \text{SVD}(W^T \Sigma_2 W)\).
  2. \( \lambda \leftarrow \max_{d, \nu}(d/\nu)\).
  3. \( \Sigma_2 \leftarrow V \text{diag}(\lambda) V^T\).

- **Exact solution only requires one SVD**
- **Time complexity:** \( O(\max(md, md^2)) \)

Results (mean squared error):

<table>
<thead>
<tr>
<th>Method</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
<th>4th</th>
<th>5th</th>
<th>6th</th>
<th>7th</th>
<th>School</th>
</tr>
</thead>
<tbody>
<tr>
<td>STL</td>
<td>31.41</td>
<td>22.90</td>
<td>9.13</td>
<td>10.30</td>
<td>0.14</td>
<td>0.84</td>
<td>0.46</td>
<td>0.9882 ± 0.0196</td>
</tr>
<tr>
<td>MTFL</td>
<td>31.41</td>
<td>22.91</td>
<td>9.13</td>
<td>10.33</td>
<td>0.14</td>
<td>0.83</td>
<td>0.45</td>
<td>0.8891 ± 0.0380</td>
</tr>
<tr>
<td>MTRL</td>
<td>31.09</td>
<td>22.69</td>
<td>9.08</td>
<td>9.74</td>
<td>0.14</td>
<td>0.83</td>
<td>0.44</td>
<td>0.9007 ± 0.0407</td>
</tr>
<tr>
<td>MTFLR</td>
<td>31.13</td>
<td>22.60</td>
<td>9.10</td>
<td>9.74</td>
<td>0.13</td>
<td>0.83</td>
<td>0.45</td>
<td>0.8451 ± 0.0197</td>
</tr>
<tr>
<td>FEATR</td>
<td>31.08</td>
<td>22.68</td>
<td>9.08</td>
<td>9.73</td>
<td>0.13</td>
<td>0.83</td>
<td>0.43</td>
<td>0.8134 ± 0.0253</td>
</tr>
<tr>
<td>STL-NM</td>
<td>24.81</td>
<td>17.20</td>
<td>8.97</td>
<td>8.36</td>
<td>0.13</td>
<td>0.72</td>
<td>0.34</td>
<td>–</td>
</tr>
<tr>
<td>MT-NM</td>
<td>12.01</td>
<td>10.54</td>
<td>5.02</td>
<td>7.15</td>
<td>0.09</td>
<td>0.70</td>
<td>0.27</td>
<td>–</td>
</tr>
<tr>
<td>MTFLR-NM</td>
<td>11.02</td>
<td>9.51</td>
<td>4.99</td>
<td>7.11</td>
<td>0.08</td>
<td>0.62</td>
<td>0.27</td>
<td>–</td>
</tr>
<tr>
<td>FEATR-NM</td>
<td>10.77</td>
<td>9.34</td>
<td>4.95</td>
<td>7.01</td>
<td>0.08</td>
<td>0.59</td>
<td>0.24</td>
<td>–</td>
</tr>
</tbody>
</table>

Feature covariance matrix and task covariance matrix:

(a) Covariance matrix over features. (b) Covariance matrix over tasks.