Low-rank Interaction with Sparse Additive Effects Model for Large Data Frames

Geneviève Robin, Hoi-To Wai, Julie Josse, Olga Klopp, Éric Moulines

December 6, 2018
Thirty-second Conference on Neural Information Processing Systems

Poster #87
5-7pm
Motivation: species monitoring

White headed duck: endangered
- lead poisoning
- wetland loss

Eurasian curlew: declining
- lead poisoning
- habitat destruction
- disturbances

### Waterbirds counts

<table>
<thead>
<tr>
<th>Site</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>site 1</td>
<td>NA</td>
<td>16</td>
<td>32</td>
</tr>
<tr>
<td>site 2</td>
<td>299</td>
<td>286</td>
<td>346</td>
</tr>
<tr>
<td>site 3</td>
<td>NA</td>
<td>96</td>
<td>151</td>
</tr>
<tr>
<td>site 4</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>site 5</td>
<td>NA</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>site 6</td>
<td>4647</td>
<td>6054</td>
<td>2442</td>
</tr>
<tr>
<td>site 7</td>
<td>16</td>
<td>45</td>
<td>30</td>
</tr>
<tr>
<td>site 8</td>
<td>5916</td>
<td>6485</td>
<td>1249</td>
</tr>
</tbody>
</table>

### Sites and year covariates

<table>
<thead>
<tr>
<th>Site</th>
<th>Surface</th>
<th>Country</th>
<th>Latitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.35</td>
<td>Algeria</td>
<td>36.64</td>
</tr>
<tr>
<td>2</td>
<td>15.4</td>
<td>Tunisia</td>
<td>34.11</td>
</tr>
<tr>
<td>3</td>
<td>1.12</td>
<td>Lybia</td>
<td>35.77</td>
</tr>
<tr>
<td>4</td>
<td>0.34</td>
<td>Morocco</td>
<td>35.51</td>
</tr>
<tr>
<td>5</td>
<td>2.8</td>
<td>Algeria</td>
<td>34.49</td>
</tr>
<tr>
<td>6</td>
<td>2.6</td>
<td>Algeria</td>
<td>35.91</td>
</tr>
<tr>
<td>7</td>
<td>0.98</td>
<td>Tunisia</td>
<td>35.75</td>
</tr>
<tr>
<td>8</td>
<td>7.2</td>
<td>Morocco</td>
<td>30.36</td>
</tr>
</tbody>
</table>

### 1) Characteristics of the data
- **Mixed**: categorical, real and discrete
- **Large scale**: 25,000+ survey sites
- **Incomplete**: missing values
- **Side information**: row & column covariates

### 2) Goal: estimate
- **Main effects**: effect of covariates
- **Interactions**: the remaining effects
Low-rank Interaction and Sparse main effects

Heterogeneous exponential family parametric model:

\[ f_{Y_{ij}}(y) = f_{ij}(y, X_{ij}) \]

Main effects and interactions in parameter space:

\[ X_{ij} = \langle u_{ij}, \alpha \rangle + L_{ij} \]

Estimation:

\[ (\hat{\alpha}, \hat{L}) \in \arg\min \mathcal{L}(Y; X) + \lambda_1 \|L\|_* + \lambda_2 \|\alpha\|_1 \]

Two-fold generalisation of “sparse plus low-rank” matrix recovery

1. general sparsity pattern
2. exponential family noise
Statistical guarantees

\[
(\hat{\alpha}, \hat{L}) \in \arg\min L(Y; X) + \lambda_1 \| L \|_\star + \lambda_2 \| \alpha \|_1
\]

Near optimal error bounds for main effects and interactions

**Theorem 1:**

\[
\| \hat{\alpha} - \alpha^0 \|_2^2 \leq \frac{\| \alpha^0 \|_1}{\pi} \times \max_k \| U(k) \|_1 + D_\alpha
\]

\[
\| \hat{L} - L^0 \|_F^2 \leq \frac{\text{rank}(L^0) \max(n, p)}{\pi} + D_L
\]

Convergence results

**Mixed Coordinate Gradient Descent Algorithm:**

- proximal update for \( \alpha \)
- conditional gradient/Franke-Wolfe update for \( L \)

Sublinear convergence and computationally efficient

**Theorem 2:**

The MCGD method converges to an \( \epsilon \)-solution in \( \mathcal{O}(1/\epsilon) \) iterations
- **Fast in large dimensions**
- **Estimation of main effects constant with dimensions**
- **Robust to large proportions of missing values**
Poster #87
5-7pm