**Polyadic Regression and its Application to Chemogenomics**

Ioakeim (Kimis) Perros¹, Fei Wang², Ping Zhang³, Peter Walker⁴, Richard Vuduc¹, Jyotishman Pathak², Jimeng Sun¹

¹Georgia Institute of Technology, ²Weill Cornell Medicine, ³IBM T.J. Watson Research Center, *Naval Medical Research Center

---

**Drug-perturbed gene expression analysis**

![Drug-perturbed gene expression analysis](image)

Lab measurement $y_{i1,i2,i3}$ indicates effectiveness of drug $i_1$ towards treating tissue $i_3$, w.r.t. gene $i_2$.

**Importance of accurate measurement estimates**

- Drug repositioning (reduced cost, shorter path to approval)
- Revealing drug mechanisms, precise identification of targets
- For some drugs, no measurements may be available. We should still enable estimation, using external knowledge

**Success story: drug repositioning**

Anti-ulcer drug cimetidine was predicted and validated to be therapeutic of lung adenocarcinoma in mice

**Contributions**

- Predicts measurements associated with multiple objects (polyadic prediction); generalizes dyadic prediction
- Can incorporate external knowledge to enable predictions for new objects, unobserved during training
- Expressive model exploring all high-order interactions between different data domains efficiently

**Polyadic Prediction use case**

![Polyadic Prediction use case](image)

**Core model**

\[ f(x_{i1}^1, x_{i2}^2, \ldots, x_{iK}^K) = b + \sum_{k=1}^{K} \alpha_k x_{i1}^k + \sum_{k=1}^{K} \beta_k x_{i2}^k + \sum_{k=1}^{K} \gamma_k x_{i3}^k + \sum_{k=1}^{K} \delta_k x_{i4}^k + \ldots + \sum_{k=1}^{K} \epsilon_k x_{i5}^k \]

\[ + \sum_{k=1}^{K} \sum_{l=1}^{K} S_{kl1} x_{i6}^k x_{i7}^l + \sum_{k=1}^{K} \sum_{l=1}^{K} S_{kl2} x_{i8}^k x_{i9}^l + \ldots + \sum_{k=1}^{K} \sum_{l=1}^{K} S_{klm} x_{i10}^k x_{i11}^l \]

General polyadic interactions

\[ S_{klm} = \mathbf{S}^{klm} \mathbf{F}^k \mathbf{F}^l \mathbf{F}^m \]

**Factorized model**

- Interactions are not independent: we pursue the low-rank structure of the matrix/tensor of feature interactions
- Shared structure exists among interactions of different order (e.g., the low-rank approximation of $\sum_{k=1}^{K} \sum_{l=1}^{K} S_{klm} x_{i6}^k x_{i7}^l x_{i8}^l$)

**Synthetic data: sparse support recovery**

- Can we identify the interacting features?
- Uniform random generation of features and low-rank factors

\[ f(x_{i1}^1, x_{i2}^2, x_{i3}^3) = (x_{i1}^1)^T \mathbf{F}^1 \mathbf{C}^{11} \mathbf{F}^1 x_{i1}^1 + (x_{i2}^2)^T \mathbf{F}^2 \mathbf{C}^{22} \mathbf{F}^2 x_{i2}^2 + (x_{i3}^3)^T \mathbf{F}^3 \mathbf{C}^{33} \mathbf{F}^3 x_{i3}^3 + (x_{i1}^1)^T \mathbf{F}^1 \mathbf{C}^{12} \mathbf{F}^2 x_{i2}^2 + \cdots + (x_{i1}^1)^T \mathbf{F}^1 \mathbf{C}^{1m} \mathbf{F}^m x_{im} + \epsilon \]

- Only 10% of features are relevant (rest of $\mathbf{F}^k$ rows set to zero)
- 100 objects per data domain (1 Mil. instances), $d_k = 50$ features, $m_k = 5$, $f_2$ regularization on $\mathbf{F}^k$

**Real data experimental setup**

- LINCS L1000 publicly-available drug-gene-tissue data: ~1000 genes, known to be maximally predictive
- Used 10 out of 56 tissues with the most expression profiles, discarded 128 genes for which no similarity information was available (850 genes considered)
- Task 1: pick the 81 drugs measured in all top-10 tissues (total of 688,500 expression values)
- Task 2: 500 drugs having measurements in at least 2 of top-10 tissues & constrain the train, validation and test sets to have no common drugs (total of 1,870,850 expression values)

**Task 1: Estimating missing measurements**

- Linear terms: low predictive value
- MMVs: joint factorization and regularization (even if linear terms are irrelevant)
- PMs: competitive, but do not include 3-order interactions
- Polyadic Regression: highest accuracy

**Task 2: Predicting measurements for new drugs**

- MMVs: joint factorization and regularization (even if linear terms are irrelevant)
- PMs: competitive, but do not include 3-order interactions
- Polyadic Regression: highest accuracy

**Scalability: Avg. running time**

<table>
<thead>
<tr>
<th>Method</th>
<th>Training time (sec)</th>
<th>Test time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polyadic Regression</td>
<td>0.2305 ± 0.006896</td>
<td>4417</td>
</tr>
<tr>
<td>Factorization Machines</td>
<td>0.1272 ± 0.005342</td>
<td>4417</td>
</tr>
<tr>
<td>Multiview Machines</td>
<td>0.0069 ± 0.001727</td>
<td>4425</td>
</tr>
</tbody>
</table>

**Reference**


E-mail: perros@gatech.edu