Probabilistic Matrix Addition

Amrudin Agovic, Arubdan Banerjee, Snigdhansu Chatterjee

Dept of Computer Science & Engg, University of Minnesota
School of Statistics, University of Minnesota

Presented by Lingbo Li

ICML, 2011
Motivation

- Most data represented in matrix form
- Protein function prediction, Geo-statistics, ...
- Data assumed independently identically distributed (i.i.d.)
- How do we move away from i.i.d assumption?
- Wish to incorporate relationships among rows and columns
- Problems of interest:
  - Matrix Completion
  - Matrix Extension
  - Matrix Completion and Extension
Probabilistic Matrix Addition (PMA)

- Let $X$ be a $n \times m$ matrix
- Assume two Gaussian Processes: $G_1 \equiv GP(0, \mathcal{K}_1)$ and $G_2 \equiv GP(0, \mathcal{K}_2)$ with covariance function $\mathcal{K}_1$ and $\mathcal{K}_2$ corresponding to rows and columns respectively
- Sample $f_1, ..., f_m$ where $f \in \mathbb{R}^n$ independently from $G_1$: $p(f|G_1) = \frac{1}{(2\pi)^{n/2}|\mathcal{K}_1|^{1/2}} \exp\left(-\frac{1}{2}f^T\mathcal{K}_1^{-1}f\right)$
- Sample $g_1, ..., g_n$ where $g \in \mathbb{R}^m$ independently from $G_2$: $p(g|G_2) = \frac{1}{(2\pi)^{m/2}|\mathcal{K}_2|^{1/2}} \exp\left(-\frac{1}{2}g^T\mathcal{K}_2^{-1}g\right)$

\[ F = \begin{bmatrix} f_1 & \cdots & f_m \end{bmatrix} = \begin{bmatrix} f_1(1) & \cdots & f_m(1) \\ \vdots & \ddots & \vdots \\ f_1(n) & \cdots & f_m(n) \end{bmatrix} \]
\[ G = \begin{bmatrix} g_1^T \\ \vdots \\ g_n^T \end{bmatrix} = \begin{bmatrix} g_1(1) & \cdots & g_1(m) \\ \vdots & \ddots & \vdots \\ g_n(1) & \cdots & g_n(m) \end{bmatrix} \]
Given the two random matrices $F$ and $G$, generate the $n \times m$ random matrix $X$ as $X = F + G$

For an entry of $X$: $x_{ij} = f_j(i) + g_i(j)$
Marginal distribution of $x_{ij}$ is a univariate Gaussian:
\[ x_{ij} \sim \mathcal{N}(0, \mathcal{K}_{1,(i,i)} + \mathcal{K}_{2,(j,j)}) \]

\[ E[f_j(i)f_j(l)] = \mathcal{K}_1(i, l), \quad E[f_j(i)f_k(l)] = 0, \]
\[ E[g_i(j)g_i(k)] = \mathcal{K}_2(j, k), \quad E[g_i(j)g_l(k)] = 0 \]

\[ E[x_{ij}^2] = \mathcal{K}_1(i, i) + \mathcal{K}_2(j, j), \quad E[x_{ij}x_{ik}] = \mathcal{K}_2(j, k), \]
\[ E[x_{ij}x_{lj}] = \mathcal{K}_1(i, l), \quad \text{and} \quad E[x_{ij}x_{lk}] = 0 \]

Consider vectorized matrix $\text{vec}(X) \in \mathbb{R}^{mn} \sim \mathcal{N}(0, \sum_{\text{vec}(X)})$
where $\sum_{\text{vec}(X)} = (\mathbb{I}_m \otimes \mathcal{K}_1) + (\mathbb{I}_n \otimes \mathcal{K}_2) = \mathcal{K}_1 \oplus \mathcal{K}_2$

Recall, in LMC we have
\[ \sum_{\text{vec}(X)} = \sum_{u} \mathcal{K}^{(u)}_1 \otimes \mathcal{K}^{(u)}_2 \]

Difference:
\[ \text{In LMCs covariances combined using Kronecker products} \]
\[ \text{PMA has sparse dependency structure} \]
PMA: Conditional Distributions

- Given $F_{(-i,-j)}$, $f_j(i)$ only depends on $f_j(-i)$, the other elements of $f_j$, and is conditionally independent of $G_{(-i,-j)}$
- Given $G_{(-i,-j)}$, $g_i(j)$ only depends on $g_i(-j)$, the other elements of $g_i$, and is conditionally independent of $F_{(-i,-j)}$

$$f_j(i) | (F_{(-i,-j)}, G_{(-i,-j)}) \sim N(m_j^f(i), s_j^f(i))$$
$$g_i(j) | (F_{(-i,-j)}, G_{(-i,-j)}) \sim N(m_i^g(j), s_i^g(j))$$

where

$$m_j^f(i) = \mathcal{K}_{1,(i,-i)}\mathcal{K}_{1,(-i,-i)}^{-1}f_j(-i),$$
$$s_j^f(i) = \mathcal{K}_{1,(i,i)} - \mathcal{K}_{1,(i,-i)}\mathcal{K}_{1,(-i,-i)}^{-1}\mathcal{K}_{1,(-i,i)},$$
$$m_i^g(j) = \mathcal{K}_{2,(j,-j)}\mathcal{K}_{2,(-j,-j)}^{-1}g_i(-j),$$
$$s_i^g(j) = \mathcal{K}_{2,(j,j)} - \mathcal{K}_{2,(j,-j)}\mathcal{K}_{2,(-j,-j)}^{-1}\mathcal{K}_{2,(-j,j)}.$$

Since $x_{ij} = f_j(i) + g_i(j)$, we have

$$x_{ij} | (F_{(-i,-j)}, G_{(-i,-j)}) \sim N(m_{ij}, s_{ij})$$

where

$$m_{ij} = m_j^f(i) + m_i^g(j), \quad s_{ij} = s_j^f(i) + s_i^g(j).$$
Possible, follows from GP regression

Consider matrix completion, \( P \) unknown values, \( N = mn \)

\[
\sum_{\text{vec}(X)} = \left[ \begin{array}{cc} \sum_{T} (N-P) \times (N-P) & \sum_{P} (N-P) \times P \\ \sum_{(N-P) \times P} & \sum_{P \times P} \end{array} \right]
\]  

For predictions we obtain

\[
X_u = \sum_{P \times (N-P)} \sum_{(N-P) \times (N-P)}^{-1} X_k
\]

\( \sum_{(N-P) \times (N-P)} \) is an arbitrary and potentially large submatrix
PMA: Approximation Inference by Gibbs Samplings

- Let \( \tilde{X} \) be a full matrix, where missing values have been initialized to random values.
- Since \( X = F + G \), only need to sample \( F \) or \( G \) for given gram matrices \((K_1, K_2)\).
- If \( K_1 \) and/or \( K_2 \) is unknown, alternate between sampling \((G, X)\) and estimating \( K_1 \) and/or \( K_2 \).
- **Sampling \( G \):**

\[
p(g_i(j)|G_{(-i,-j)}, \tilde{X}, K_1, K_2) \propto p(g_i(j)|G_{(-i,-j), X_{(i,-j)}, K_1, K_2})p(x_{ij}|G, X_{(-i,-j), K_1, K_2})
\]

\[
= p(g_i(j)|G_{(-i,-j), K_2})p(\tilde{x}_{ij} - g_i(j)|F_{(-i,-j), K_1})
\]

\[
\propto \mathcal{N}(g_i(j)|\mu_{ij}, \sigma_{ij}^2)
\]

where \( \mu_{ij} = \frac{m^g_i(j)s^f_j(i)+m^f_i(j)s^g_j(j)}{s^f_j(i)+s^g_j(j)} \), and \( \sigma_{ij}^2 = \frac{s^f_j(i)s^g_j(j)}{s^f_j(i)+s^g_j(j)} \).

- Inversion kept efficient by using rank-two updates.
Sampling $X$: for missing value prediction, sample missing values $p(\tilde{x}_{ij}|\tilde{X}_{(-i,-j)}, F, K_1, K_2) = \mathcal{N}(x_{ij}|\bar{x}_{ij}, \zeta_{ij})$ where

$$\bar{x}_{ij} = f_j(i) + K_{2,(i,-i)}K_{2,(-i,-i)}^{-1}(\tilde{x}_{-i,j} - f_j(-i)),$$

$$\zeta_{ij} = K_{2,(i,i)} - K_{2,(i,-i)}K_{2,(-i,-i)}^{-1}K_{2,(-i,i)}$$

Parameter Estimation: If $K_1$ and $K_2$ are unknown, we initialize $\hat{K}_1 \succ 0, \hat{K}_2 \succ 0$, and alternate between sampling $(G, \tilde{X})$ and estimating $(\hat{K}_1, \hat{K}_2)$. We have already outlined how to sample $G$ and $\tilde{X}$. Let $F = \tilde{X} - G$. Then, we have $\hat{K}_1 = \frac{1}{m} \sum_{i=1}^{m} f_j f_j^T$ and $\hat{K}_2 = \frac{1}{n} \sum_{i=1}^{n} g_i g_i^T$. 
PMA: MAP Inference

- **Estimating** $F$: the joint log-likelihood over $(X, F)$ is:
  \[
  \log p(\tilde{X}, F|K_1, K_2) = \log p(F|K_1) + \sum_{i=1}^{n} \log p(\tilde{x}_i|f:(i), K_2)
  \]

- Optimization satisfies the Sylvester equation
  \[
  FK_2 + K_1 F = K_1 \hat{X}
  \]

- Estimating $\hat{X}$:
  \[
  \hat{x}_{ij}^{\text{new}} = f_{j}(i) + K_{2,(j,-j)K_{2,(-j,-j)}}^{-1}(\hat{x}_{-i,j} - f_{j}(-i))
  \]

*Parameter Estimation:* We initialize $\hat{K}_1 > 0, \hat{K}_2 > 0$, and the missing values of $X$ randomly. Then, we alternate between updating $(\tilde{X}, F)$, and estimating $\hat{K}_1$, $\hat{K}_2$. We have already discussed updates for $(\tilde{X}, F)$. Let $G = \tilde{X} - F$. Then $\hat{K}_1 = \frac{1}{m} \sum_{j=1}^{m} f_j f_j^T$ and $\hat{K}_2 = \frac{1}{n} \sum_{i=1}^{n} g_i g_i^T$. 
PMA: Matrix Completion results on artificial data

- Simulated data generated using PMA model, of size $50 \times 20$ and $50 \times 50$
- Use mean square error of the predicted values for evaluation
- Compare PMA to GP regression across rows or columns
- Modeling correlations across both rows and columns is better than ones across either rows or columns

(a) 50x20 Matrix  
(b) 50x50 Matrix
PMA: Matrix Completion Results on benchmark datasets

- Binary entries translated to truncated log odds
- PMA does not suffer by assuming a sparser dependency structure
- PMA-GIBBS performs slightly better than PMA-MAP
- PMA is competitive compared to PMF

Table 1. Error rates for recovering missing labels obtained using five-fold cross validation on the Emotions and Scene datasets. Performance of GPR, PMA-MAP, PMA-GIBBS, PMA-EXACT, I-LMC and PMF is evaluated while an increasing percentage of labels are missing in the training data. Missing Labels are randomly selected. The error rates reflect the percentage of missing labels incorrectly recovered.

<table>
<thead>
<tr>
<th></th>
<th>10%</th>
<th>15%</th>
<th>20%</th>
<th>25%</th>
<th>30%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Emotions</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPR</td>
<td>32.3 ± 5.9</td>
<td>33.1 ± 4.7</td>
<td>32.6 ± 5.0</td>
<td>34.6 ± 2.3</td>
<td>35.3 ± 2.3</td>
</tr>
<tr>
<td>PMA-MAP</td>
<td>23.9 ± 6.5</td>
<td>25.3 ± 3.8</td>
<td>26.9 ± 4.4</td>
<td>29.7 ± 4.8</td>
<td>30.8 ± 4.7</td>
</tr>
<tr>
<td>PMA-GIBBS</td>
<td>23.3 ± 5.3</td>
<td>24.8 ± 3.2</td>
<td>25.1 ± 3.8</td>
<td>27.2 ± 3.9</td>
<td>28.0 ± 4.0</td>
</tr>
<tr>
<td>PMA-EXACT</td>
<td>19.7 ± 4.9</td>
<td>23.6 ± 6.9</td>
<td>25.8 ± 4.0</td>
<td>27.3 ± 5.4</td>
<td>27.9 ± 4.1</td>
</tr>
<tr>
<td>I-LMC</td>
<td>20.3 ± 4.6</td>
<td>25.1 ± 5.9</td>
<td>25.7 ± 3.7</td>
<td>27.6 ± 4.5</td>
<td>27.8 ± 3.8</td>
</tr>
<tr>
<td>PMF</td>
<td>21.8 ± 5.0</td>
<td>22.6 ± 2.4</td>
<td>24.6 ± 3.0</td>
<td>26.3 ± 1.6</td>
<td>26.0 ± 3.7</td>
</tr>
<tr>
<td><strong>Scene</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GPR</td>
<td>14.7 ± 1.7</td>
<td>34.5 ± 8.0</td>
<td>17.2 ± 2.1</td>
<td>17.4 ± 1.7</td>
<td>18.0 ± 2.1</td>
</tr>
<tr>
<td>PMA-MAP</td>
<td>11.9 ± 1.0</td>
<td>13.6 ± 2.5</td>
<td>13.8 ± 2.7</td>
<td>13.9 ± 3.2</td>
<td>14.8 ± 1.5</td>
</tr>
<tr>
<td>PMA-GIBBS</td>
<td>10.3 ± 1.4</td>
<td>10.9 ± 2.6</td>
<td>11.1 ± 1.8</td>
<td>11.3 ± 2.1</td>
<td>12.3 ± 1.2</td>
</tr>
<tr>
<td>PMA-EXACT</td>
<td>10.4 ± 1.0</td>
<td>11.0 ± 1.0</td>
<td>11.6 ± 1.8</td>
<td>11.9 ± 1.2</td>
<td>12.5 ± 2.3</td>
</tr>
<tr>
<td>I-LMC</td>
<td>10.4 ± 1.0</td>
<td>10.9 ± 1.0</td>
<td>11.8 ± 1.7</td>
<td>11.8 ± 1.2</td>
<td>12.9 ± 2.6</td>
</tr>
<tr>
<td>PMF</td>
<td><strong>9.2 ± 2.2</strong></td>
<td>13.8 ± 3.0</td>
<td>16.1 ± 3.4</td>
<td>18.5 ± 2.8</td>
<td>20.1 ± 3.0</td>
</tr>
</tbody>
</table>
Predicting new rows

- Predict a new row in the data matrix $X$ assuming that $K_1$ is known kernel function
- First initialize the new row $x_{(n+1)}$ of $X$
  - Use GP regression to obtain estimate for new row $f_j(n+1)$ yielding the extended matrix $\tilde{F} \in \mathbb{R}^{(n+1) \times m}$:
    \[
p(f_\ast|f) = \frac{1}{(2\pi)^{1/2}|\tilde{K}|^{1/2}} \exp\left(-\frac{1}{2}(f_\ast - \bar{f})^T \tilde{K}^{-1} (f_\ast - \bar{f})\right)
    \]
    where $\tilde{K} = K_{\ast,\ast} - K_{\ast,f}K_{f,f}^{-1}K_{f,\ast}$, and $\bar{f} = K_{\ast,f}K_{f,f}^{-1}f$
  - sample a new row $g_{n+1}^T \sim GP(0, K_2)$ yielding the extended $\tilde{G} \in \mathbb{R}^{(n+1) \times m}$
- For Gibbs Sampling, obtain the initial extended matrix as $\tilde{X} = \tilde{F} + \tilde{G}$, then proceed Gibbs Sampling inference, treating the entire last row of $\tilde{X}$ as latent
- For MAP inference, the $(n + 1)^{st}$ row of $\tilde{F}$ serves directly as an estimate for the new row $x_{(n+1)}$:
- If $K_2$ is unknown, alternate sampling/estimating $\tilde{X}$ and estimating $K_2$
PMA: Matrix Extension

- $\mathcal{K}_1$ assumed to be RBF kernel, $\mathcal{K}_2$ estimated
- Use the Scene and Emotions datasets
- Use truncated log-odds for learning, and the sign of the predicted score for evaluation
- Evaluate PMA-GIBBS with three multi-label classification algorithms
PMA: Matrix Completion and Extension

- RBF kernel across rows, label covariance estimated
- Partially labeled training data
- Improvement most visible

Table 2. Five fold cross validation on the Scene data set with 25% of label entries missing. PMA-GIBBS utilizes all available data while training. PMA-GIBBS-D, MLKNN-D, IBLRML-D and MLSVM-D discard data points with missing label entries in the training stage.

<table>
<thead>
<tr>
<th>Model</th>
<th>OneError</th>
<th>AvePrec</th>
<th>Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>PMA-GIBBS</td>
<td>29.7 ± 4.2</td>
<td>82.3 ± 2.7</td>
<td>10.6 ± 2.3</td>
</tr>
<tr>
<td>PMA-GIBBS-D</td>
<td>51.1 ± 7.5</td>
<td>67.5 ± 4.8</td>
<td>22.4 ± 3.7</td>
</tr>
<tr>
<td>MLKNN-D</td>
<td>70.5 ± 2.0</td>
<td>46.3 ± 4.3</td>
<td>23.7 ± 8.1</td>
</tr>
<tr>
<td>IBLRML-D</td>
<td>61.9 ± 8.9</td>
<td>36.9 ± 3.9</td>
<td>54.8 ± 3.9</td>
</tr>
<tr>
<td>MLSVM-D</td>
<td>87.9 ± 2.0</td>
<td>40.9 ± 1.6</td>
<td>83.1 ± 1.4</td>
</tr>
</tbody>
</table>
Conclusions

▶ Introduced PMA, a model for arbitrary real-valued matrices
▶ Moves away from i.i.d. assumption
▶ Does not factorize over rows or columns
▶ Utilizes sparse dependency structure
▶ Competitive performance to state-of-the-art

1

1Original slides for ICML presentation(http://techtalks.tv/talks/54427/) are referred