Non-Gaussian Discriminative Factor Models via the Max-Margin Rank-Likelihood

Presented by Wenzhao Lian

Xin Yuan @ Bell Labs, Alcatel-Lucent, Ricardo Henao, Ephraim L. Tsalik, Raymond Langley, and Lawrence Carin @ Duke University

July 8, 2015
**Intuition**

**Figure**: Top and right panels are group-wise empirical distributions for \( \text{rank}(x) \) and \( \log(x) \), respectively.

Both log transformation and the rank can be used for classification.
Challenge and Solution

- Challenge: Data from computational biology and bioinformatics are rarely Gaussian and often discrete.
- Solution: Directly model the rank of the data.
- Benefits:
  1) Treat ordinal, continuous and discrete data within the same framework.
  2) Make weaker assumptions about the distribution of the data.
Ordinal probit model

Consider the ordinal data $d \times N$ matrix $X$,

$$x_{i,n} = g_i(w_{i,n}),$$  \hspace{1cm} (1)

$$w_{i,n} = a_i^\top z_n + v_{i,n},$$  \hspace{1cm} (2)

$$v_{i,n} \sim \mathcal{N}(0, 1).$$  \hspace{1cm} (3)

- $A = [a_1 \ldots a_d]^\top \in \mathbb{R}^{d \times K}$, $Z \in \mathbb{R}^{K \times N}$, $W \in \mathbb{R}^{d \times N}$.
- $g_i(\cdot)$ is a non-decreasing function such that the rankings of the $N$ realizations of component $i$ are preserved. Large values in $x_i$ (rows of $X$) correspond to large values in $w_i$ (rows of $W$).

$$x_{i,n} = g_i(w_{i,n}) = j \quad \text{if} \quad h_{i,j-1} < w_{i,n} < h_{i,j};$$  \hspace{1cm} (4)

$h_{i,0} = -\infty$, $h_{i,J_i} = \infty$ and $h_i = [h_{i,1} \ldots h_{i,J_i-1}]$ is a vector of thresholds for the $i$th row of $W$. 
Rank-likelihood model

Provided that \( g_i(w_{i,n}) \) is non-decreasing, we know that given \( x_{i,n} < x_{i,n'} \), then \( g_i(w_{i,n}) < g_i(w_{i,n'}) \) and \( w_{i,n} < w_{i,n'} \), thus

\[
R(x_i) = \{ w_i \in \mathbb{R}^N : w_{i,n} < w_{i,n'} \text{ if } x_{i,n} < x_{i,n'} \} \quad (5)
\]

\( R(x_i) \) is the set of all possible vectors \( w_i \) such that

\[
\text{rank}(x_i) = \text{rank}(w_i) \quad (6)
\]

Joint probability distribution of the model

\[
\prod_{i=1}^{d} p(w_i \in R(x_i), a_i, Z) = p(A)p(Z) \prod_{i=1}^{d} \left\{ \int_{R(x_i)} \prod_{n=1}^{N} \mathcal{N}(w_{i,n}; a_i^\top z_n, 1) \, dw_{i,n} \right\} \quad (7)
\]
• Inference:

\[ p(w_{i,n} | w_{i\setminus n}, a_i, z_n) = p(w_{i,n} | w_{i,l}^i, w_{i,u}^i, a_i, z_n) \]
\[ = T \mathcal{N}(a_i^\top z_n, 1, w_{i,l}^i, w_{i,u}^i) \] (8)

\[ w_{i,l}^i = \max\{w_{i,n'} : x_i,n' < x_i,n\} \]
\[ w_{i,u}^i = \min\{w_{i,n'} : x_i,n < x_i,n'\} \]

• Disadvantage:
Numerically unstable and computationally expensive.

• Solution:
The max-margin rank-likelihood
Max-margin rank-likelihood

\[ R_{mm}(x_i) = \{ w_i \in \mathbb{R}^N : w_{i,n} < w_{i,n'} - \epsilon \text{ if } x_{i,n} < x_{i,n'} \} \]

(9)

Any two distinct elements of \( w_i \) must be separated by a gap of size no smaller than \( \epsilon > 0 \).

\[
\max\{ w_{i,n'} : x_{i,n'} < x_{i,n} \} + \epsilon < w_{i,n} < \min\{ w_{i,n'} : x_{i,n} < x_{i,n'} \} - \epsilon
\]

Figure: Graphical representation of the loss function associated to the max-margin rank-likelihood, where \( \ell^u_\epsilon + \ell^l_\epsilon = \ell_\epsilon(w_{i,n} - w^u_{i,n}) + \ell_\epsilon(w^l_{i,n} - w_{i,n}) \). Note that \( w^l_{i,n} + \epsilon < w_{i,n} < w^u_{i,n} - \epsilon \) is not penalized by the loss function.
Max-margin rank-likelihood

A pseudo-likelihood for $w_{i,n}$ is

$$L_i(w_{i,n}|w_{i\setminus n}) = \exp \left\{ -\ell_\epsilon(w_{i,n} - w_{i,n}^u) - \ell_\epsilon(w_{i,n}^l - w_{i,n}) \right\}$$  \hspace{1cm} (10)

$$w_{i,n} = a_i^\top z_n, \quad w_{i,n}^u = a_i^\top z_n^u, \quad w_{i,n}^l = a_i^\top z_n^l$$  \hspace{1cm} (11)

- $\ell_\epsilon(u) = 2 \max(0, u + \epsilon)$ can be interpreted as the one-sided $\epsilon$-sensitive loss.
- Maximizing (10) is equivalent to finding $w_i \in R_{\text{mm}}(x_i)$ such that differences between neighbor elements of $w_i$ are maximal given $\epsilon$, hence the term max-margin is used.
Polson & Scott has shown $\ell_\epsilon(u) = 2 \max(0, u + \epsilon)$ admits a location-scale mixture of Gaussians,

$$\exp\{-2 \max(0, u)\} = \int \mathcal{N}(u; -\lambda, \lambda) d\lambda:$$

$$L_i(w_{i,n} | w_{i\setminus n}) = \int \mathcal{N}(w_{i,n} - w_{i,u}; -\epsilon - \lambda_{i,n}^u, \lambda_{i,n}^u) \times \mathcal{N}(w_{i,n} - w_{i,l}; -\epsilon - \lambda_{i,n}^l, \lambda_{i,n}^l) d\lambda_{i,n}^u d\lambda_{i,n}^l \quad (12)$$

Data augmentation via latent variables $\{\lambda_{i,n}^u\}$ and $\{\lambda_{i,n}^l\}$. 

Non-Gaussian Discriminative Factor Models via the Max-Margin Rank-Likelihood July 8, 2015
Bayesian SVM based discriminative factor model

Introducing labels $y = [y_1 \ldots y_N]^\top \in \{-1, 1\}$, using the Bayesian SVM:

$$L_n(y_n|\beta, z_n) = \exp\{-2 \max(0, u_n)\}$$

$$= \int_{0}^{\infty} \frac{1}{\sqrt{2\pi} \lambda_n^c} \exp \left(-\frac{1}{2} \frac{(u_n + \lambda_n^c)^2}{\lambda_n^c} \right) d\lambda_n^c$$  \hspace{1cm} (13)

$$u_n = 1 - y_n \beta^\top z_n, \ \beta \in \mathbb{R}^K$$ is a vector of classifier coefficients and $\{\lambda_n^c\}$ is a vector of latent variables, with superscript $c$ denoting the classifier.

Linear discriminative model via three-parameter-beta normal (TPBN) priors:

Imposing sparsity on the factor loadings $\mathbf{A}$ and classifier weights $\mathbf{\beta}$

$$a_{i,k} \sim \text{TPBN}(r_a, s_a, \Phi_{k}^{(a)}), \quad z_n \sim \mathcal{N}(0, I_K) \quad \beta_k \sim \text{TPBN}(r_\beta, s_\beta, \Phi^{(\beta)})$$

Specifically:

$$a_{i,k} \sim \mathcal{N}(0, \xi_{i,k}), \quad \xi_{i,k} \sim \text{Ga}(r_a, \eta_{i,k}), \quad \eta_{i,k} \sim \text{Ga}(s_a, \Phi_{k}^{(a)}),$$

$$\beta_k \sim \mathcal{N}(0, b_k), \quad b_k \sim \text{Ga}(r_\beta, e_k), \quad e_k \sim \text{Ga}(s_\beta, \Phi^{(\beta)})$$

$$\Phi_{k}^{(a)} \sim \text{Ga}(0.5, \tilde{\Phi}), \quad \tilde{\Phi} \sim \text{Ga}(0.5, 1),$$

$$\Phi^{(\beta)} \sim \text{Ga}(0.5, \bar{\Phi}), \quad \bar{\Phi} \sim \text{Ga}(0.5, 1).$$
Non-Linear discriminative model via DP

Build a **global nonlinear** decision rule as a mixture of **local linear** classifiers:

\[ G = \sum_{t=1}^{\infty} q_t \delta_{\theta^*_t}, \quad q_t = \nu_t \prod_{l=1}^{t-1} (1 - \nu_l), \quad \nu_t \sim \text{Beta}(1, \alpha), \quad \theta^*_t \sim G_0, \]

where \(\sum_{t=1}^{\infty} q_t = 1\), \(\delta_{\theta^*_t}\) represents a point measure at \(\theta^*_t\) and \(\alpha\) is the concentration parameter.

\[ y_n, z_n \sim f(\theta_n), \quad \theta_n \sim G \quad (14) \]

\[ f(\theta_n) = \ln(y_n|\beta_n, z_n) \mathcal{N}(z_n|\mu_n, \psi_n^{-1}I_K) \]

\[ G_0 = T \mathcal{P}BN(\beta|r_\beta, s_\beta, \Phi(\beta)) \times \mathcal{N}(\mu|0, I_K) \times \text{Ga}(\psi|\psi_s, \psi_r) \]

\(\{\beta_n, \mu_n, \psi_n\} = \{\beta_t, \mu_t, \psi_t\}\) if sample \(n\) belongs to the \(t\)-th component.
Inference and Prediction

- Inference: both MCMC and VB have been derived.
- Prediction: we use the pair \( \{y, X\} \) to estimate the parameters of the model (training), namely \( \{A, z, \beta\} \), then given a test point \( x_* \), we go through three steps:
  
  (i) Compare \( x_* \) to \( X \) to determine the rank of each component of \( x_* \) w.r.t. to the training data, which amounts to finding \( \{w_{i,*}^l, w_{i,*}^u\} \), for \( i = 1, \ldots, d \).

  (ii) For fixed \( \{A, w_{i,*}^l, w_{i,*}^u\} \), estimate \( z_* \) from its conditional posterior.

  (iii) Make a prediction for \( x_* \) using \( \text{sign}(\beta^\top z_*) \).
Model compositions and handwritten digits results

**Table**: Composition of different methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Likelihood</th>
<th>Classifier</th>
<th>DPM</th>
</tr>
</thead>
<tbody>
<tr>
<td>G-L-Probit</td>
<td>Gaussian</td>
<td>probit</td>
<td>No</td>
</tr>
<tr>
<td>G-L-BSVM</td>
<td>Gaussian</td>
<td>BSVM</td>
<td>No</td>
</tr>
<tr>
<td>OR-L-Probit</td>
<td>ordinary rank</td>
<td>probit</td>
<td>No</td>
</tr>
<tr>
<td>OR-L-BSVM</td>
<td>ordinary rank</td>
<td>BSVM</td>
<td>No</td>
</tr>
<tr>
<td>R-L-BSVM</td>
<td>max-margin rank</td>
<td>BSVM</td>
<td>No</td>
</tr>
<tr>
<td>G-NL-BSVM</td>
<td>Gaussian</td>
<td>BSVM</td>
<td>Yes</td>
</tr>
<tr>
<td>R-NL-BSVM</td>
<td>max-margin rank</td>
<td>BSVM</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Table**: Mean error rates (%) and runtime in seconds for USPS 3 vs. 5 subtask.

<table>
<thead>
<tr>
<th></th>
<th>G-L-Probit</th>
<th>G-L-BSVM</th>
<th>OR-L-Probit</th>
<th>OR-L-BSVM</th>
<th>R-L-BSVM</th>
<th>G-NL-BSVM</th>
<th>R-NL-BSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>5.95±0.005</td>
<td>5.86±0.008</td>
<td>5.05±0.013</td>
<td>4.92±0.027</td>
<td>4.53±0.026</td>
<td>3.88±0.017</td>
<td>3.23±0.025</td>
</tr>
<tr>
<td>Time</td>
<td>8.64</td>
<td>10.29</td>
<td>14.07</td>
<td>14.19</td>
<td>16.05</td>
<td>23.81</td>
<td>36.63</td>
</tr>
</tbody>
</table>

**Table**: Mean error rates (%) and runtime in seconds for MNIST 3 vs. 5 subtask.

<table>
<thead>
<tr>
<th></th>
<th>G-L-BSVM</th>
<th>R-L-BSVM</th>
<th>L-SVM</th>
<th>G-NL-BSVM</th>
<th>R-NL-BSVM</th>
<th>NL-SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>5.05±0.053</td>
<td>4.84±0.014</td>
<td>4.68</td>
<td>4.21±0.010</td>
<td>2.10±0.007</td>
<td>2.00</td>
</tr>
<tr>
<td>Runtime</td>
<td>150</td>
<td>220</td>
<td>140</td>
<td>400</td>
<td>600</td>
<td>304</td>
</tr>
</tbody>
</table>

Non-Gaussian Discriminative Factor Models via the Max-Margin Rank-Likelihood  July 8, 2015 14
Gene expression: data

A tuberculosis data (GEO accession series GSE39941)

- 47323 genes and 334 subjects.
- Four phenotypes: active tuberculosis (TB) (111), latent TB (54), TB of both kinds (165), HIV positive (107).
- Keep the top 4732 genes with largest intensity profiles.
Gene expression: results

Table 5: AUC (with error bars), accuracy and runtime in seconds for gene expression data.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PFA-L-BSVM</th>
<th>G-L-BSVM</th>
<th>R-L-BSVM</th>
<th>G-NL-BSVM</th>
<th>R-NL-BSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>TB vs. Others</td>
<td>0.740±0.102, 0.683</td>
<td>0.766±0.093, 0.704</td>
<td>0.814±0.052, 0.740</td>
<td>0.847±0.061, 0.778</td>
<td>0.872±0.025, 0.781</td>
</tr>
<tr>
<td>Active TB vs. Others</td>
<td>0.802±0.070, 0.775</td>
<td>0.857±0.050, 0.784</td>
<td>0.896±0.028, 0.832</td>
<td>0.921±0.034, 0.853</td>
<td>0.948±0.021, 0.880</td>
</tr>
<tr>
<td>Latent TB vs. Others</td>
<td>0.849±0.051, 0.802</td>
<td>0.907±0.037, 0.841</td>
<td>0.923±0.041, 0.868</td>
<td>0.934±0.029, 0.874</td>
<td>0.954±0.025, 0.889</td>
</tr>
<tr>
<td>HIV(+) vs. HIV(-)</td>
<td>0.850±0.056, 0.793</td>
<td>0.879±0.055, 0.844</td>
<td>0.900±0.055, 0.856</td>
<td>0.915±0.041, 0.850</td>
<td>0.959±0.051, 0.901</td>
</tr>
<tr>
<td>One fold time</td>
<td>130</td>
<td>141</td>
<td>180</td>
<td>330</td>
<td>450</td>
</tr>
</tbody>
</table>

Figure: The learned coefficients $\beta$ for the 4 classifiers (one vs. others) based on gene expression data.
RNA sequencing

A RNA sequencing sepsis data:

- 133 subjects and 15158 genes
- three different groups, systemic inflammatory response (SIRS) (26), sepsis survivors (SeS) (78) and sepsis complications leading to death (SeD) (29).

<table>
<thead>
<tr>
<th>Methods</th>
<th>PFA-L-BSVM</th>
<th>G-L-BSVM</th>
<th>R-L-BSVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRS vs. Se</td>
<td>0.70±0.04, 0.73</td>
<td>0.78±0.02, 0.76</td>
<td>0.86±0.01, 0.81</td>
</tr>
<tr>
<td>SeD vs. SeS</td>
<td>0.76±0.05, 0.70</td>
<td>0.76±0.01, 0.75</td>
<td>0.82±0.02, 0.78</td>
</tr>
<tr>
<td>SIRS vs. SeS</td>
<td>0.75±0.02, 0.71</td>
<td>0.87±0.01, 0.71</td>
<td>0.91±0.01, 0.87</td>
</tr>
<tr>
<td>One fold time</td>
<td>179</td>
<td>175</td>
<td>226</td>
</tr>
</tbody>
</table>
(1) We introduce a new max-margin version of the rank-likelihood model towards Bayesian factor analysis.

(2) We propose a discriminative factor model by integrating max-margin rank-likelihood.

(3) We extend the discriminative factor model to nonlinear decision functions, through a mixture of local linear classifiers.
Q & A

xyuan@bell-labs.com
Bell Labs, Alcatel-Lucent