1 Stirling Numbers

The Stirling numbers of the three kinds are three different ways to partition \( y \) elements into \( n \) groups.

- The Stirling number of the first kind corresponds to the number of ways of partitioning \( y \) elements into \( n \) disjoints cycles.
- The Stirling number of the second kind corresponds to the number of ways of partitioning \( y \) elements into \( n \) non-empty subsets.
- The Stirling number of the third kind (also known as Lah number) corresponds to the number of ways of partitioning \( y \) elements into \( n \) non-empty ordered subsets.

\[
\begin{array}{c|c|c}
\text{First kind} & \text{Second kind} & \text{Third kind} \\
St_1(3,1) = 2 & St_2(3,1) = 1 & St_3(3,1) = 6 \\
\end{array}
\]

Figure 1: Illustration of the Stirling numbers of the three kinds for \( y = 3 \) and \( n = 1 \).

2 Proof of limit cases

**Proposition 1.** If there exists \( \theta^{\text{raw}} \) such that
\[
\lim_{\theta \to \theta^{\text{raw}}} \kappa^T \psi(\theta) = -\infty,
\]
then the posterior of dcPF tends to the posterior of PF as \( \theta \) goes to \( \theta^{\text{raw}} \).

**Proposition 2.** If there exists \( \theta^{\text{bin}} \) such that
\[
\lim_{\theta \to \theta^{\text{bin}}} \kappa^T \psi(\theta) = +\infty,
\]
then the posterior of dcPF tends to the posterior of PF applied to binarized data as \( \theta \) goes to \( \theta^{\text{bin}} \), i.e.: \[
\lim_{\theta \to \theta^{\text{bin}}} p(W, H | Y) = p(W, H | N = Y^b).
\]

**Proof.** Let \( \lambda \in \mathbb{R}_+ \), \( n \sim \text{Poisson}(\lambda) \) and \( y|n \sim ED(\theta, n\kappa) \) with support given by \( S = \{n, \ldots, +\infty\} \):

\[
p(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!},
\]

(1)

\[
p(y|n) = \exp(y\theta - n\kappa^T \psi(\theta)) h(y, n\kappa), \quad y \in S,
\]

(2)

where \( \kappa \) and \( \psi(\theta) \) can either be scalars or vectors of the same dimension. In both cases, \( \kappa^T \psi(\theta) \in \mathbb{R} \). We denote by \( r = \lambda e^{-\kappa^T \psi(\theta)} \).

We have the following posterior distribution for \( y > 0 \):

\[
p(n|y) = \frac{r^n h(y, n\kappa) (n!)^{-1}}{\sum_{m=1}^{y} r^m h(y, m\kappa) (m!)^{-1}}, \quad n \in \{1, \ldots, y\}.
\]

(3)

Thus, for fixed \( \kappa \) and \( y > 0 \), we have that:

\[
\sum_{m=1}^{y} r^m h(y, m\kappa) (m!)^{-1} \sim_{r \to +\infty} r^y h(y, y\kappa) (y!)^{-1}
\]

(4)

\[
\sim_{r \to 0} r h(y, \kappa).
\]

(5)

It follows:

\[
p(n|y) \xrightarrow{r \to +\infty} \delta_y(n)
\]

(6)

\[
p(n|y) \xrightarrow{r \to 0} \delta_1(n).
\]

(7)

From these results we can deduce that, in dcPF, assuming:

- there exists \( \theta^{\text{raw}} \) such that \( \lim_{\theta \to \theta^{\text{raw}}} \kappa^T \psi(\theta) = -\infty \),
- there exists \( \theta^{\text{bin}} \) such that \( \lim_{\theta \to \theta^{\text{bin}}} \kappa^T \psi(\theta) = +\infty \).
Then, we have the following limit cases:

\[
p(N | Y) = \int_{W,H} p(N | Y, W, H)p(W, H | Y)dWdH
\]
\[
\lim_{\theta \to \theta^{\text{raw}}} \int_{W,H} \delta_Y(N) p(W, H | Y)dWdH = \delta_Y(N)
\]
\[
\lim_{\theta \to \theta^{\text{bin}}} \int_{W,H} \delta_{Y^b}(N) p(W, H | Y)dWdH = \delta_{Y^b}(N).
\]

(8)

And finally, for the posterior distribution:

\[
p(W, H | Y) = \int_{N} p(W, H | N)p(N | Y) dN
\]
\[
\lim_{\theta \to \theta^{\text{raw}}} p(W, H | N = Y)
\]
\[
\lim_{\theta \to \theta^{\text{bin}}} p(W, H | N = Y^b),
\]

(9) (10) (11)

where \(p(W, H | N)\) is the posterior of a PF model with raw or binarized observations respectively.

3 Adaptivity of dcPF to over-dispersion

Table 1: Mean, variance and ratio var/mean of the non-zero values for each dataset. Learned parameters for each model and each dataset.

<table>
<thead>
<tr>
<th></th>
<th>Taste Profile</th>
<th>NIPS</th>
<th>Last.fm</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean of non-zeros</td>
<td>2.66</td>
<td>2.74</td>
<td>3.86</td>
</tr>
<tr>
<td>var of non-zeros</td>
<td>25.94</td>
<td>20.87</td>
<td>65.72</td>
</tr>
<tr>
<td>ratio var/mean</td>
<td>9.8</td>
<td>7.6</td>
<td>17.0</td>
</tr>
<tr>
<td>Log - (p)</td>
<td>0.80</td>
<td>0.74</td>
<td>0.90</td>
</tr>
<tr>
<td>ZTP - (p)</td>
<td>1.95</td>
<td>1.40</td>
<td>2.35</td>
</tr>
<tr>
<td>Geo - (p)</td>
<td>0.60</td>
<td>0.51</td>
<td>0.69</td>
</tr>
<tr>
<td>sh. NB - (p)</td>
<td>0.87</td>
<td>0.86</td>
<td>0.90</td>
</tr>
<tr>
<td>sh. NB - (\kappa_2)</td>
<td>0.21</td>
<td>0.17</td>
<td>0.27</td>
</tr>
</tbody>
</table>

Table 1 illustrates how the natural parameter \(\theta = \log(p)\) is strongly correlated to the variance-mean ratio of the non-zero values of the datasets. Hence, it illustrates the adaptivity of dcPF to over-dispersion.