Supplementary Material: Recommendation from Raw Data with Adaptive Compound Poisson Factorization

Olivier Gouvert, Thomas Oberlin, Cédric Févotte
IRIT, Université de Toulouse, CNRS, France
firstname.lastname@irit.fr

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.

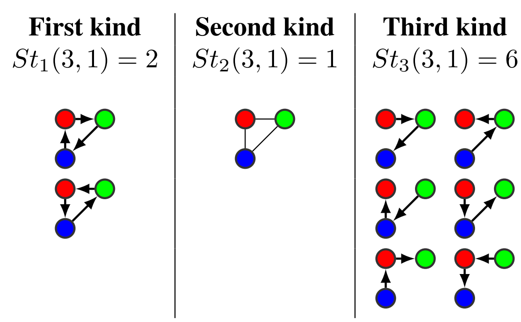

Figure 1: Illustration of the Stirling numbers of the three kinds for $y = 3$ and $n = 1$.

2 Proof of limit cases

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

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

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

$$p(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!}, \tag{1}$$

$$p(y|n) = \exp(y\theta - n\kappa^T \psi(\theta)) h(y, n\kappa), \; y \in S, \tag{2}$$

where κ 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 $y > 0$:

$$p(n|y) = \frac{r^n h(y, n\kappa)(n!)^{-1}}{\sum_{m=1}^{y} r^m h(y, m\kappa)(m!)^{-1}}, \; n \in \{1, \ldots, y\}. \tag{3}$$

Thus, for fixed κ and $y > 0$, we have that:

$$\sum_{m=1}^{y} r^m h(y, m\kappa)(m!)^{-1} \underset{r \to +\infty}{\sim} r^y h(y, y\kappa)(y!)^{-1} \tag{4}$$

$$\underset{r \to 0}{\sim} r h(y, \kappa). \tag{5}$$

It follows:

$$p(n|y) \xrightarrow[r \to +\infty]{} \delta_y(n) \tag{6}$$

$$p(n|y) \xrightarrow[r \to 0]{} \delta_1(n). \tag{7}$$

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

• there exists θ^{raw} such that $\lim_{\theta \to \theta^{\text{raw}}} \kappa^T \psi(\theta) = -\infty$,

• there exists θ^{bin} such that $\lim_{\theta \to \theta^{\text{bin}}} \kappa^T \psi(\theta) = +\infty$.

Then, we have the following limit cases:

$$p(\mathbf{N}|\mathbf{Y}) = \int\limits_{\mathbf{W},\mathbf{H}} p(\mathbf{N}|\mathbf{Y},\mathbf{W},\mathbf{H}) p(\mathbf{W},\mathbf{H}|\mathbf{Y}) d\mathbf{W} d\mathbf{H}$$
$$\xrightarrow[\theta \to \theta^{\text{raw}}]{} \int\limits_{\mathbf{W},\mathbf{H}} \delta_{\mathbf{Y}}(\mathbf{N})\, p(\mathbf{W},\mathbf{H}|\mathbf{Y}) d\mathbf{W} d\mathbf{H} = \delta_{\mathbf{Y}}(\mathbf{N})$$
$$\xrightarrow[\theta \to \theta^{\text{bin}}]{} \int\limits_{\mathbf{W},\mathbf{H}} \delta_{\mathbf{Y}^b}(\mathbf{N})\, p(\mathbf{W},\mathbf{H}|\mathbf{Y}) d\mathbf{W} d\mathbf{H} = \delta_{\mathbf{Y}^b}(\mathbf{N}). \tag{8}$$

And finally, for the posterior distribution:

$$p(\mathbf{W},\mathbf{H}|\mathbf{Y}) = \int\limits_{\mathbf{N}} p(\mathbf{W},\mathbf{H}|\mathbf{N}) p(\mathbf{N}|\mathbf{Y}) d\mathbf{N} \tag{9}$$
$$\xrightarrow[\theta \to \theta^{\text{raw}}]{} p(\mathbf{W},\mathbf{H}|\mathbf{N} = \mathbf{Y}) \tag{10}$$
$$\xrightarrow[\theta \to \theta^{\text{bin}}]{} p(\mathbf{W},\mathbf{H}|\mathbf{N} = \mathbf{Y}^b), \tag{11}$$

where $p(\mathbf{W},\mathbf{H}|\mathbf{N})$ 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.

	Taste Profile	**NIPS**	**Last.fm**
mean of non-zeros	2.66	2.74	3.86
var of non-zeros	25.94	20.87	65.72
ratio var/mean	9.8	7.6	17.0
Log - p	0.80	0.74	0.90
ZTP - p	1.95	1.40	2.35
Geo - p	0.60	0.51	0.69
sh. NB - p	0.87	0.86	0.90
sh. NB - κ_2	0.21	0.17	0.27

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.