Discrete Adduct Clustering Model for MS Data

February 6, 2015

1 Binning

We can bin the data along the mass dimension. Specifically, given N peak features in a file, we create a corresponding N bins centered at the precursor mass M by applying the inverse M+H adduct transformation from the peak mass O

$$M = \frac{O|c| + ce - \sum_i h_i G_i}{n} \tag{1}$$

where c is the charge, e is the mass of the electron, h_i and G_i are the adduct parts. Further details can be found in Adduct_notes.pdf.

Given the n-th precursor mass M_n, we then create a mass bin such that $M_n \pm b_n$, where $b_n = M_n * tol * (1e - 6)$ with the value of tol specified by the user. Repeat this for all N precursor masses created for each peak, so we end up with K bins, where $N = K$. Each bin now define a valid precursor mass cluster that a peak can be assigned to based on the potential adduct transformation. We index peak features by $n = 1, ..., N$ and precursor mass bin (i.e. mass clusters) by $k = 1, ..., K$.

2 Model

Denote the peak feature by $d_n = (x_n, y_n)$ where x_n is the mass value and y_n the RT value. We use the variable $z_n = k$ to denote the assignment of peak feature n to bin k.

Given the data, we want to infer the assignment of the z_n variables to the clusters. Assume a fixed number of clusters (based on the known 'valid' precursor masses $K = N$). Each categorical variables $z_1, ..., z_n$ is then independently drawn from a categorical distribution with parameter $\boldsymbol{\theta}$ and determines the assignment of peak n to cluster k. The parameter vector $\boldsymbol{\theta}$ of length K is drawn from a Dirichlet distribution with parameter α. The likelihood of a peak into a cluster also depends on whether there's a possible transformation from its ion mass to the precursor mass and based on the RT value too.

$$\begin{aligned} \boldsymbol{\theta} &\sim Dir(\alpha) && (2) \\ z_n = k &\sim Cat(\boldsymbol{\theta}) && (3) \\ d_n &\sim L(d_n|z_n = k, ...) && (4) \end{aligned}$$

The likelihood $L(d_n|z_n = k, ...)$ can be factorised into the mass and RT terms $L(d_n|z_n = k) = p(x_n|z_n = k) \cdot p(x_y|z_n = k)$. For the mass term,

$$p(x_n|z_n = k) = I_k(x_n) \tag{5}$$

where $I_k(x_n)$ is the indicator function that produces 1 if there is a possible adduct inverse transformation from ion mass x_n to bin k and 0 otherwise. y_n is normally distributed with mean equals to μ_k, the RT value of the peak that produce the mass bin during the binning stage, and some variance σ_k^2.

$$p(y_n|z_n = k) = \mathcal{N}(y_n|\mu_k, \sigma_k^2) \tag{6}$$

The assumption implicit in the model is that a peak must always be assignable to a mass bin.

For Gibbs sampling,

$$\begin{aligned} p(z_n = k|\boldsymbol{\theta}, ...) &\propto (\alpha_k + z_k) \cdot L(d_n|z_n = k) && (7) \\ &= (\alpha_k + z_k) \cdot I_k(x_n) \cdot \mathcal{N}(y_n|\mu_k, \sigma_k^2) && (8) \end{aligned}$$

Full derivations to follow in below section.

2.1 Derivations

Some standard derivations for Dirichlet-discrete model.

The joint distribution of the data is

$$\begin{aligned} p(z_1, ..., z_n, \boldsymbol{\theta}, \alpha) &= \prod_n p(z_n = k|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\alpha) && (9) \\ p(z_n = k|\boldsymbol{\theta}) &= Cat(\boldsymbol{\theta}) = \prod_k (\boldsymbol{\theta}_k)^{z_k} && (10) \\ p(\boldsymbol{\theta}|\alpha) &= Dir(\alpha) = \frac{1}{B(\alpha)} \prod_k (\boldsymbol{\theta}_k)^{\alpha_k - 1} && (11) \end{aligned}$$

For Gibbs sampling, we need $p(z_n|\boldsymbol{\theta}, ...)$. This is

$$\begin{aligned} p(z_n = k|\boldsymbol{\theta}, ...) &\propto p(z_n = k, \boldsymbol{\theta}, ...) && (12) \\ &= p(z_n = k|\boldsymbol{\theta}, ...) \cdot p(\boldsymbol{\theta}|\alpha) && (13) \end{aligned}$$

$$
\begin{aligned}
&= \int \left[p(z_n = k|\boldsymbol{\theta}, ...) \cdot p(\boldsymbol{\theta}|\alpha)\right] \, d\theta && (14)\\
&= \int \left[\prod_k (\boldsymbol{\theta}_k)^{z_k} \cdot \frac{1}{B(\alpha)} \prod_k (\boldsymbol{\theta}_k)^{\alpha_k - 1}\right] d\theta && (15)\\
&= \frac{1}{B(\alpha)} \int \left[\prod_k (\boldsymbol{\theta}_k)^{z_k + \alpha_k - 1}\right] d\theta && (16)\\
&= \frac{B(\alpha + z)}{B(\alpha)} && (17)
\end{aligned}
$$

Eq (16) is obtained as such: we know

$$
B(\alpha) = \int \left[\prod_k (\boldsymbol{\theta}_k)^{z_k + \alpha_k - 1}\right] d\theta \tag{18}
$$

as it's the normalising constant in the multinomial pdf, so

$$
B(\alpha + z) = \int \left[\prod_k (\boldsymbol{\theta}_k)^{z_k + \alpha_k - 1}\right] d\theta \tag{19}
$$

Continuing from eq (17), the beta function is defined as $B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$. Expand everything, and we get

$$
\begin{aligned}
p(z_n = k|\boldsymbol{\theta}, ...) &\propto \left(\frac{\prod_k \Gamma(\alpha_k + z_k)}{\Gamma(\sum_k \alpha_k + z_k)}\right)\left(\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)}\right) && (20)\\
&= \left(\frac{\Gamma(\sum_k \alpha_k)}{\Gamma(\sum_k \alpha_k + z_k)}\right)\left(\frac{\prod_k \Gamma(\alpha_k + z_k)}{\prod_k \Gamma(\alpha_k)}\right) && (21)\\
&\propto \prod_k \frac{\Gamma(\alpha_k + z_k)}{\Gamma(\alpha_k)} && (22)\\
&\propto \prod_k \Gamma(\alpha_k + z_k) && (23)\\
&\propto \text{...?} && (24)\\
&\propto \alpha_k + z_k && (25)
\end{aligned}
$$

References