Implementation of Bayesian Hierachical Clustering

Lina Yang, Xiaodi Qin

April 23, 2017

1 Abstract

from paper (Heller and Ghahramani, 2005)

2 Background

3 Algorithm

4 Optimization

5 Application

5.1 Simulated data sets

5.2 Real data sets

6 Comparative analysis

7 Discussion

8 Code

The repository can be found at `https://github.com/qxxxd/Bayesian-Hierarchical-Clustering`.

Algorithm 1: Bayesian Hierachical Clustering

```
Input: Data $X = (X_0, X_1, ..., X_N)$, $family \in \{niw\}$, hyperparameter $\alpha$, scaling factor on
       the prior precision of the mean $r$.
Output: A linkage matrix $Z$
For $l$ in RANGE($N$):
    $n_l = 1$, $d_l = \alpha$
    $ml_l = family(X_l)$
t = 0
For $i$ in RANGE($N - 1$):
    For $j$ in RANGE($i + 1, N$):
        $c1_t = i$, $c2_t = j$, $n_t = n_i + n_j$, $d_t = \alpha\Gamma(n_t) + d_i d_j$
        $\pi_t = \alpha\Gamma(n_t)/d_t$
        $X_t = (X_i, X_j)^T$
        $\mathrm{P}(D_t, H_1^t) = family(X_t)\pi_t$, $\mathrm{P}(D_t, H_2^t) = ml_i \times ml_j(d_i d_j/d_t)$
        $logodds_t = \log \mathrm{P}(D_t, H_1^t) - \log \mathrm{P}(D_t, H_2^t)$
        $t = t + 1$
$rm = []$, $Z = []$
For $p$ in RANGE($N$-$1$):
    $idx = \arg\max_{\{idx \in \{0,...,t\}, c1_{idx} \notin rm, c2_{idx} \notin rm\}} logodds$
    $Z$.APPEND($[c1_{idx}, c2_{idx}, logodds_{idx}, n_{idx}]$)
    $rm$.APPEND($c1_{idx}, c2_{idx}$)
    $maxlogodds = -Inf$
    For $q$ in RANGE($t$):
        If $c1_q \notin rm$ and $c2_q \notin rm$:
            $c1_{temp} = N + p$, $c2_{temp} = q$, $n_{temp} = n_{idx} + n_q$, $d_{temp} = \alpha\Gamma(n_{temp}) + d_{idx} d_q$
            $\pi_{temp} = \alpha\Gamma(n_{temp})/d_{temp}$
            $X_{temp} = (X_{idx}, X_q)^T$, $ml_{idx} = family(X_{idx})$, $ml_q = family(X_q)$
            $\mathrm{P}(D_{temp}, H_1^{temp}) = family(X_{temp})\pi_t$,
             $\mathrm{P}(D_{temp}, H_2^{temp}) = ml_{idx} \times ml_q(d_{idx} d_q/d_{temp})$
            $logodds_{temp} = \log \mathrm{P}(D_{temp}, H_1^{temp}) - \log \mathrm{P}(D_{temp}, H_2^{temp})$
            If $logodds_{temp} > maxlogodds$:
                $c1_t = c1_{temp}, c2_t = c2_{temp}, n_t = n_{temp}$
                $d_t = d_{temp}, logodds_t = logodds_{temp}$
        t = t + 1
return $Z$
```

9 References

Heller, K. A. and Z. Ghahramani (2005). Bayesian Hierarchical Clustering. In *Proceedings of the 22Nd International Conference on Machine Learning*, ICML '05, New York, NY, USA, pp. 297–304. ACM.