Supplementary material for "A Biclustering Approach Based on Factor Graphs and The Max-Sum Algorithm"

M. Denitto, A. Farinelli, M.A.T. Figueiredo, M. Bicego

1 Messages derivation

In this section the derivations of each message update rule are presented. Given a general variable x and a general factor f each message is derived following the Max-Sum rules:

1. from factors to variables:

$$\mu_{f\to x}(x) = \max_{x_1 \ldots x_n} \left[f(x, x_1, \ldots, x_n) + \sum_{m \in ne(f)\setminus x} \mu_{x_m \to f}(x_m) \right] \tag{1}$$

2. from variables to factors:

$$\mu_{x\to f}(x) = \sum_{l \in ne(x)\setminus f} \mu_{f_l \to x}(x). \tag{2}$$

Note that a general message $\mu(x)$ should be a vector storing the message value for each possible configuration of x. For instance, if x can take k possible values μ should be a vector $\mu = [\mu(x = 1), \cdots, \mu(x = k)]$. Once Max-Sum converges the value of a variable is assigned as the value for whom the sum of the incoming messages is maximum. Specifically in the case of binary variables we should memorize the value of $\mu(x = 0)$ and $\mu(x = 1)$. However, without loss of information, we can store the difference between this values (i.e. $\mu = \mu(x = 1) - \mu(x = 0)$) and assign 1 to a variable if the message is positive and 0 otherwise. The following messages are derived on the basis of this consideration.

Notation: given a data matrix A with $N = \{1, \ldots, n\}$ set of rows and $M = \{1, \ldots, m\}$ set of columns, in what follows $i, t \in N$ and $t, k \in M$. And $\hat{i} \in N\setminus i$, $\hat{k} \in M\setminus k$ and $\hat{tk} = \{(1,1), \cdots, (n,m)\}\setminus(t,k)$

Summary The Factor Graph we propose is composed by binary variables c_{ij} stating if a point belongs to the solution ($c_{ij} = 1$) or not ($c_{ij} = 0$). The factors of the

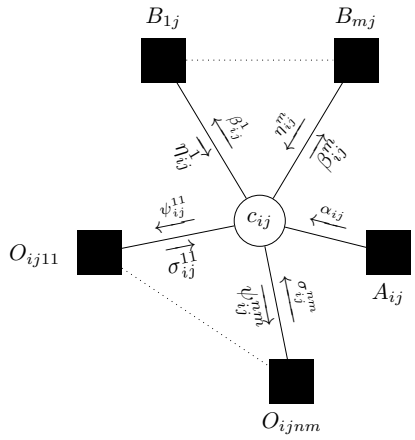

Figure 1: Sketch of the factor graph showing the connections between one variable and all its factors

proposed Factor Graphs are:

$$A_{ij}(c_{ij}) = \begin{cases} a_{i,j} & \text{if } c_{ij} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

$$O_{a_{ij},a_{tk}}(c_{ij}, c_{tk}) = \begin{cases} w * I(a_{ij}, a_{tk}) & \text{if } c_{ij} = c_{tk} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

$$B_{jk}(c_{1j}, \ldots, c_{nk}) = \begin{cases} 0 & \begin{array}{l}\text{if } \sum_i c_{ij} = 0 \\ \text{or } \sum_i c_{ik} = 0 \\ \text{or } \sum_i (c_{ij} - c_{ik}) = 0\end{array} \\ -\infty & \text{otherwise} \end{cases} \tag{5}$$

Resulting in the Factor Graph sketched in figure 1.

1.1 Messages from Variables to Factors

Following Equation (2) the value of the message going from a variable x to factor f is the sum of the incoming messages without considering the one arriving from f.

Given the Factor Graph in Figure 1 we obtain:

- $\psi_{ij}^{tk}(c_{ij}) = \mu_{c_{ij} \to O_{ijtk}}(c_{ij})$
  - if $c_{ij} = 1$

$$\psi_{ij}^{tk}(1) = \sum_{\hat{tk}} \sigma_{ij}^{\hat{tk}}(1) + \sum_{k} \eta_{ij}^{k}(1) + \alpha_{ij}(1)$$

– if $c_{ij} = 0$

$$\psi_{ij}^{tk}(0) = \sum_{\hat{tk}} \sigma_{ij}^{\hat{tk}}(0) + \sum_{k} \eta_{ij}^{k}(0) + \alpha_{ij}(0)$$

– then

$$\begin{aligned}\psi_{ij}^{tk} &= \psi_{ij}^{tk}(c_{ij})(1) - \psi_{ij}^{tk}(c_{ij})(0) \\ &= \sum_{\hat{tk}} \sigma_{ij}^{\hat{tk}} + \sum_{k} \eta_{ij}^{k} + \alpha_{ij}\end{aligned}$$

- $\beta_{ij}^{k}(c_{ij}) = \mu_{c_{ij} \to B_{jk}}(c_{ij})$
  – if $c_{ij} = 1$

$$\beta_{ij}^{k}(1) = \sum_{\hat{k}} \eta_{ij}^{\hat{k}}(1) + \sum_{tk} \sigma_{ij}^{tk}(1) + \alpha_{ij}(1)$$

  – if $c_{ij} = 0$

$$\beta_{ij}^{k}(0) = \sum_{\hat{k}} \eta_{ij}^{\hat{k}}(1) + \sum_{tk} \sigma_{ij}^{tk}(0) + \alpha_{ij}(0)$$

  – then

$$\begin{aligned}\beta_{ij}^{k} &= \beta_{ij}^{k}(1) - \beta_{ij}^{k}(0) \\ &= \sum_{\hat{k}} \eta_{ij}^{\hat{k}} + \sum_{tk} \sigma_{ij}^{tk} + \alpha_{ij}\end{aligned}$$

- The message going from a variable c_{ij} to its function A_{ij} is not named in Figure 1 because it is commonly set to 0, and then not considered[Frey and Dueck(2007)]. The intuition behind this choice is that the information provided by A_{ij} (i.e. the entry value) must not change among the Max-Sum iterations.

1.2 Messages from Factors to Variables

These derivations represent the bottleneck of many approaches. In fact the maximization in Equation (1) is intractable since, for a given function and considering integer variables, we should analyse d^k configurations (with d variables domain size and k number of neighbours). In this section we present how we can exploit the binary nature of the variables and the hard constraints to reduce the possible configurations in such maximization.

Note: we exploit the following properties in the derivation:

Rule 1) $a - \max(a, b) = \min(0, a - b)$;
Rule 2) $\max(a - b) - b = \max(a - b, 0)$;
Rule 3) $\max(a, b) - \max(c, d) = \max[\min(a - c, a - d), \min(b - c, b - d)]$;
Rule 4) $a - \max(b, c) = \min(a - b, a - c)$

Given the Factor Graph in Figure 1 we obtain:

- $\sigma_{ij}^{tk}(c_{ij}) = \mu_{O_{ijtk} \to c_{ij}}(c_{ij})$

$$\sigma_{ij}^{tk}(c_{ij}) = \max_{c_{tk}} \left[O_{ijtk}(c_{ij}, c_{tk}) + \psi_{tk}^{ij}(c_{tk}) \right]$$

  - if $c_{ij} = 1$

$$\sigma_{ij}^{tk}(1) = \max_{c_{tk}} \left[O_{ijtk}(1, c_{tk}) + \psi_{tk}^{ij}(c_{tk}) \right]$$

    we must consider two cases:

    1. $c_{tk} = 1$

$$\sigma_{ij}^{tk}(1) = w * I(a_{ij}, a_{tk}) + \psi_{tk}^{ij}(1)$$

    2. $c_{tk} = 0$

$$\sigma_{ij}^{tk}(1) = \psi_{tk}^{ij}(0)$$

  - if $c_{ij} = 0$

$$\begin{aligned} \sigma_{ij}^{tk}(0) &= \max_{c_{tk}} \left[O_{ijtk}(0, c_{tk}) + \psi_{ij}^{tk}(c_{tk}) \right] \\ &= \max_{c_{tk}} \left[\psi_{tk}^{ij}(c_{tk}) \right] \end{aligned}$$

  - then considering the equations just retrieved (i.e. the ones for $c_{ij} = 1$ and the one for $c_{ij} = 0$) we need to compute

$$\begin{aligned} \sigma_{ij}^{tk} &= \max \sigma_{ij}^{tk}(1) - \max \sigma_{ij}^{tk}(0) \\ &= \max \left[w * I(a_{ij}, a_{tk}) + \psi_{ij}^{tk}(1), \psi_{ij}^{tk}(0) \right] - \max \left[\psi_{tk}^{ij}(1), \psi_{tk}^{ij}(c_{tk}0) \right] \\ &= \max \left[\min \left(w * I(a_{ij}, a_{tk}), w * I(a_{ij}, a_{tk}) + \psi_{tk}^{ij} \right), \min \left(\psi_{tk}^{ij}, 0 \right) \right] \end{aligned}$$

- $\eta_{ij}^{k}(c_{ij}) = \mu_{B_{jk} \to c_{ij}}(c_{ij})$

$$\eta_{ij}^{k}(c_{ij}) = \max_{c_{\hat{i}j}} \left[B_{jk}(c_{1j}, \ldots, c_{ij}, \ldots, c_{nk}) + \sum_{\hat{i}} \beta_{\hat{i}j}^{k}(c_{\hat{i}j}) + \sum_{t} \beta_{tk}^{j}(c_{tk}) \right] \tag{6}$$

  - if $c_{ij} = 1$

$$\eta_{ij}^{k}(c_{ij}) = \max_{c_{\hat{i}j}} \left[B_{jk}(c_{1j}, \ldots, c_{ij} = 1, \ldots, c_{nk}) + \sum_{\hat{i}} \beta_{\hat{i}j}^{k}(c_{\hat{i}j}) + \sum_{t}^{n} \beta_{tk}^{j}(c_{tk}) \right]$$

    Surely the maximizer of such function is a configuration respecting the biclustering constraint (otherwise the B factor provides a minus infinity in the objective function). Hence, considering the constraint in Equation (5) if $c_{ij} = 1$ there are only two possible situations where the constraint is satisfied: i) the k^{th} column is completely equal to the j^{th} or ii) the k^{th} column is completely zero. In order to distinguish these quantities we introduce different notations obtaining:

1. the two columns are equal:

$$\eta_{ij}^k(1) = \max_{c_{\hat{i}j}} \left[\beta_{ik}^j(1) + \sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \beta_{\hat{i}k}^j(c_{\hat{i}j}) \right] = \eta_{ij}^k(1)^\dagger$$

note that in the second part of this equation $c_{\hat{i}j}$ is used in both betas (instead of $c_{\hat{i}k}$), this is to enforce the fact that we are considering columns with the same configuration.

2. the k^{th} column is completely zero

$$\eta_{ij}^k(1) = \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \sum_{t} \beta_{tk}^j(0) \right] = \eta_{ij}^k(1)^\ddagger$$

– if $c_{ij} = 0$

$$\eta_{ij}^k(c_{ij}) = \max_{c_{\hat{i}j}} \left[B_{jk}(c_{1j}, \ldots, c_{ij} = 0, \ldots, c_{nk}) + \sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \sum_{t}^{n} \beta_{tk}^j(c_{tk}) \right].$$

Similarly, if $c_{ij} = 0$ there are only two possible situations where the constraint is satisfied: i) the k^{th} column is completely equal to the j^{th} or ii) the j^{th} column is completely zero. Hence we obtain

1. the j^{th} column is completely zero:

$$\eta_{ij}^k(0) = \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{\hat{i}j}^k(0) + \sum_{t} \beta_{tj}^k(c_{tk}) \right] = \eta_{ij}^k(0)^\ddagger$$

2. the two columns are equal:

$$\begin{aligned} \eta_{ij}^k(0) &= \max_{c_{\hat{i}j}} \left[\beta_{ik}^j(0) + \beta_{tj}^k(1) + \beta_{tk}^j(1) + \sum_{l \in N \setminus \{i,t\}} \left(\beta_{lj}^k(c_{lj}) + \beta_{lk}^j(c_{lj}) \right) \right] \\ &= \eta_{ij}^k(0)^\dagger \end{aligned}$$

note that if the column j is not completely zero (previous case) hence there is at least another $c_{tj} = 1$ (with $t \neq i$).

– then considering the four equations just retrieved (i.e the two for $c_{ij} = 1$ and the two for $c_{ij} = 0$) we must calculate

$$\begin{aligned} \eta_{ij}^k &= \max \eta_{ij}^k(1) - \max \eta_{ij}^k(0) \\ &= \max(\eta_{ij}^k(1)^\dagger, \eta_{ij}^k(1)^\ddagger) - max(\eta_{ij}^k(0)^\ddagger, \eta_{ij}^k(0)^\dagger) \\ &= \max \Big[\min \left(\eta_{ij}^k(1)^\dagger - \eta_{ij}^k(0)^\ddagger, \eta_{ij}^k(1)^\dagger - \eta_{ij}^k(0)^\dagger \right), \\ &\qquad \min \left(\eta_{ij}^k(1)^\ddagger - \eta_{ij}^k(0)^\ddagger, \eta_{ij}^k(1)^\ddagger - \eta_{ij}^k(0)^\dagger \right) \Big] \end{aligned} \tag{7}$$

In what follows we analyse each of the terms involved in (7) separately. Referring to Section 3.3 we obtain:

$\mathbf{\Theta} = \eta_{ij}^k(1)^\dagger - \eta_{ij}^k(0)^\ddagger =$

$$= \max_{c_{\hat{i}j}} \left[\beta_{ik}^j(1) + \sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \beta_{\hat{i}k}^j(c_{\hat{i}j}) \right] - \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{\hat{i}j}^k(0) + \sum_t \beta_{tj}^k(c_{tk}) \right] =$$

$$= \beta_{ik}^j(1) - \max \beta_{ik}^j(c_{ik}) + \sum_{\hat{i}} \left[\max \left(\beta_{\hat{i}j}^k(c_{\hat{i}j}) + \beta_{ik}^j(c_{\hat{i}j}) \right) - \max \left(\beta_{\hat{i}j}^k(0) + \beta_{ik}^j(c_{\hat{i}j}) \right) \right] =$$

$$\mathbf{\Theta} = \min(0, \beta_{ik}^j) + \sum_{\hat{i}} \max \left[\min(\beta_{\hat{i}j}^k, \beta_{\hat{i}j}^k + \beta_{\hat{i}k}^j), \min(-\beta_{\hat{i}k}^j, 0) \right]$$

$\mathbf{I} = \eta_{ij}^k(1)^\dagger - \eta_{ij}^k(0)^\dagger =$

$$= \max_{c_{\hat{i}j}} \left[\beta_{ik}^j(1) + \sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \beta_{\hat{i}k}^j(c_{\hat{i}j}) \right] - + \max_{c_{\hat{i}j}} \left[\beta_{ik}^j(0) + \beta_{tj}^k(1) + \beta_{tk}^j(1) + \sum_{l \in N \setminus \{i,t\}} \left(\beta_{lj}^k(c_{lj}) + \beta_{lk}^j(c_{lj}) \right) \right] =$$

Supposing that the optimal t is given, we can write

$$= \beta_{ik}^j(1) + \beta_{tj}^k(c_{tj}) + \beta_{tk}^j(c_{tj}) - \beta_{tj}^k(1) - \beta_{tk}^j(1) + + \sum_{l \in N \setminus \{i,t\}} \left[\max \left(\beta_{lj}^k(c_{lj}) + \beta_{lk}^j(c_{lj}) \right) - \max \left(\beta_{lj}^k(c_{lj}) + \beta_{lk}^j(c_{lj}) \right) \right]$$

$$= \beta_{ik}^j + \max(0, -\beta_{tj}^k - \beta_{tk}^j).$$

Since this is the result between the difference of two maximum, and the first maximum is fixed, we can now obtain the general value for the best t by minimizing the second maximum and hence

$$\mathbf{I} = \beta_{ik}^j + \min_{\hat{i}} \left[\max(0, -\beta_{\hat{i}j}^k - \beta_{\hat{i}k}^j) \right].$$

$\mathbf{K} = \eta_{ij}^k(1)^\ddagger - \eta_{ij}^k(0)^\ddagger =$

$$= \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{\hat{i}j}^k(c_{\hat{i}j}) + \sum_t \beta_{tk}^j(0) \right] - \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{\hat{i}j}^k(0) + \sum_t \beta_{tj}^k(c_{tk}) \right] =$$

$$= \sum_{\hat{i}} \left[\max \left(\beta_{\hat{i}j}^k(c_{\hat{i}j}) + \beta_{hatik}^j(0) \right) - \max \left(\beta_{\hat{i}j}^k(0) + \beta_{hatik}^j(c_{\hat{i}k}) \right) \right] + + \beta_{ik}^j(0) - max \left(\beta_{ik}^j(0) \right)$$

the first part of this equation (the one included in the sum) can be solved adopting the third rule previously described and the final result is:

$$\mathbf{K} = \min(0, -\beta_{ik}^{j}) + \sum_{\hat{i}} \max\left[\min\left(0, -\beta_{\hat{i}k}^{j}\right), \min\left(\beta_{\hat{i}j}^{k}, \beta_{ij}^{k} - \beta_{\hat{i}k}^{j}\right)\right]$$

$\mathbf{\Lambda} = \eta_{ij}^{k}(1)^{\ddagger} - \eta_{ij}^{k}(0)^{\dagger}$ =

$$= \max_{c_{\hat{i}j}} \left[\sum_{\hat{i}} \beta_{ij}^{k}(c_{\hat{i}j}) + \sum_{t} \beta_{tk}^{j}(0)\right] -$$

$$+ \max_{c_{\hat{i}j}} \left[\beta_{ik}^{j}(0) + \beta_{tj}^{k}(1) + \beta_{tk}^{j}(1) + \sum_{l \in N \setminus \{i,t\}} \left(\beta_{lj}^{k}(c_{lj}) + \beta_{lk}^{j}(c_{lj})\right)\right]$$

As previously mentioned, also in this case we assume that the best t is given obtaining

$$= \sum_{\hat{i}} \max\left[\beta_{\hat{i}j^k}(c_{\hat{i}j}) + \beta_{\hat{i}k}^{j}(0)\right] - \beta_{tj}^{k}(1) - \beta_{tk}^{j}(1) +$$

$$- \sum_{l \in N \setminus \{i,t\}} \left[\beta_{\hat{i}j}^{k}(c_{\hat{i}j}) + \beta_{\hat{i}k}^{j}(c_{\hat{i}j})\right] =$$

$$= -\beta_{tj}^{k}(1) - \beta_{tk}^{j}(1) + \max\left(\beta_{tk}^{j}(0) + \beta_{tj}^{k}(c_{tj})\right) +$$

$$+ \sum_{l \in N \setminus \{i,t\}} \left[\max\left(\beta_{\hat{i}j}^{k}(c_{\hat{i}j}) + \beta_{\hat{i}k}^{j}(0)\right) - \max\left(\beta_{\hat{i}j}^{k}(c_{\hat{i}j}) + \beta_{\hat{i}k}^{j}(c_{\hat{i}j})\right)\right] =$$

Again, the part included in the sum can be solved adopting the third rule previously described and obtaining:

$$= \max\left(-\beta_{tk}^{j}, -\beta_{tk}^{j} - \beta_{tj}^{k}\right) +$$

$$+ \sum_{\hat{i} \neq t,i} \left[\max\left(\min(0, -\beta_{\hat{i}j}^{k} - \beta_{\hat{i}k}^{j}), \min(\beta_{\hat{i}j}^{k}, -\beta_{\hat{i}k}^{j})\right)\right]$$

This is the result for a fixed t, to obtain the general update rule we need to minimize with respect to t resulting in:

$$\mathbf{\Lambda} = \min_{t \neq i} \Big[\max\left(-\beta_{tk}^{j}, -\beta_{tk}^{j} - \beta_{tj}^{k}\right) +$$

$$+ \sum_{l \neq t,i} \left(\max\left(\min(0, -\beta_{lj}^{k} - \beta_{lk}^{j}), \min(\beta_{lj}^{k}, -\beta_{lk}^{j})\right)\right)\Big].$$

Hence the message update rule for η_{ij}^{k} is

$$\eta_{ij}^{k} = \max\left[\min(\Theta, I), \min(K, \Lambda)\right]$$

References

[Frey and Dueck(2007)] B. Frey, D. Dueck, Clustering by passing messages between data points, Science 315 (2007) 972–976.