RESEARCH ARTICLE

The Penalized Biclustering Model And Related Algorithms Supplemental Online Material

Thierry Chekouo and Alejandro Murua

Département de mathématiques et de statistique, Université de Montréal,
CP 6128, succ. centre-ville, Montréal, Québec H3C 3J7 Canada

(Received 00 Month 200x; in final form 00 Month 200x)

A. Introduction

In these sections we provide further details on the EM and Bayesian algorithms described in the main body of the paper. Section B provides the EM updating equations for the plaid model of Section 2. Section C displays the full conditional distributions of the labels and parameters associated with the penalized plaid model of Section 4. The procedure to initialize the parameters and labels for the Markov chain Monte Carlo implementation of the penalized plaid model is described in Section D. Section E gives the URL addresses of the biclustering packages used in this study, including our own package implementing the penalized plaid model.

B. The EM updating equations

Note that the bicluster and combination bicluster probabilities $p(\rho_i * \kappa_j)$ are constants depending only on the combination bicluster. We will denote them as π_k , $k = 0, 1, \dots, 2^K - 1$. Observe that $p(\rho, \kappa | Y, \bar{\theta}) = \prod_{i,j} p(\rho_i, \kappa_j | y_{ij}, \bar{\theta})$. It is straightforward to verify that

$$p(\rho_i, \kappa_j | y_{ij}, \bar{\theta}) = p(\rho_i * \kappa_j | y_{ij}, \bar{\theta}) = \frac{\frac{\pi_{cb(k)}}{\sigma(\rho_i * \kappa_j)} \phi\left((y_{ij} - \mu(\rho_i * \kappa_j, \bar{\theta}))/\sigma(\rho_i * \kappa_j)\right)}{\sum_{\rho'_i * \kappa'_j} \frac{\pi_{cb(k')}}{\sigma(\rho'_i * \kappa'_j)} \phi\left((y_{ij} - \mu(\rho'_i * \kappa'_j, \bar{\theta}))/\sigma(\rho'_i * \kappa'_j)\right)},$$

where $cb(k)$ and $cb(k')$ denote the combination biclusters associated to (ρ_i, κ_j) and (ρ'_i, κ'_j) , respectively. The maximizer of $Q(\theta | \bar{\theta})$ for the plaid model is obtained by

taking the derivatives with respect to θ . These yield:

$$\begin{aligned}\mu_k &= \frac{1}{\sum_{i,j} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_{i,j} \left\{ E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \sum_{k' \neq k} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_{k'} + \alpha_{ik'} + \beta_{jk'}) \right\} \\ \alpha_{ik} &= \frac{1}{\sum_j E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_j \left\{ E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \sum_{k' \neq k} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_{k'} + \alpha_{ik'} + \beta_{jk'}) \right\} - \mu_k \\ \beta_{jk} &= \frac{1}{\sum_i E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_i \left\{ E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \sum_{k' \neq k} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_{k'} + \alpha_{ik'} + \beta_{jk'}) \right\} - \mu_k \\ \pi_{cb(k)} &\propto \sum_{i,j} p(\rho_i * \kappa_j = \rho_{cb(k)} * \kappa_{cb(k)} | y_{ij}, \bar{\theta}),\end{aligned}$$

where $\rho_{cb(k)} * \kappa_{cb(k)}$ denotes the corresponding k -th combination bicluster, and

$$\begin{aligned}\sigma^2 &= \frac{1}{qp} \sum_{i,j} E_{\bar{\theta}} \left(y_{ij} - \sum_k \rho_{ik}\kappa_{jk} (\mu_k + \alpha_{ik} + \beta_{jk}) \right)^2 \\ &= \frac{1}{qp} \sum_{i,j} \left\{ y_{ij}^2 - 2 \sum_k E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} (\mu_k + \alpha_{ik} + \beta_{jk}) \right. \\ &\quad \left. + \sum_{k,k'} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_k + \alpha_{ik} + \beta_{jk}) (\mu_{k'} + \alpha_{ik'} + \beta_{jk'}) \right\}.\end{aligned}$$

Note that the updating equations are recursive. The parameters can be estimated using a Gauss-Seidel relaxation scheme over $k = 1, \dots, K$. For example, let the superscript “ $(t+1)$ ” denote the coefficients recently updated, and the superscript “ (t) ”, the coefficients not yet updated. Then in order to solve the system, say for α_{ik} ’s, we iterate for k within the EM iterations

$$\begin{aligned}\alpha_{ik}^{(t+1)} &= \frac{1}{\sum_j E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_j \left\{ E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \sum_{k' < k} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_{k'}^{(t+1)} + \alpha_{ik'}^{(t+1)} \right. \\ &\quad \left. + \beta_{jk'}^{(t+1)}) - \sum_{k' > k} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}\rho_{ik'}\kappa_{jk'}) (\mu_{k'}^{(t)} + \alpha_{ik'}^{(t)} + \beta_{jk'}^{(t)}) \right\} - \mu_k^{(t+1)}.\end{aligned}$$

Also note that the expectation is intractable if the number of biclusters K is large. For example, one needs to compute $E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) = \sum_{\rho_i * \kappa_j = 1} p(\rho_i * \kappa_j | y_{ij}, \bar{\theta})$, which is a sum involving 2^K terms. In the non-overlapping bicluster model, the sum reduces to one term $E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) = p(\rho_i \kappa_j = 1, \prod_{k' \neq k} (1 - \rho_{ik'} \kappa_{jk'}) = 1 | y_{ij}, \bar{\theta})$. In this latter

case, the updating equations simplify to

$$\begin{aligned}\mu_k &= \frac{1}{\sum_{i,j} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_{i,j} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} \\ \alpha_{ik} &= \frac{1}{\sum_j E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_j E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \mu_k \\ \beta_{jk} &= \frac{1}{\sum_i E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_i E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) y_{ij} - \mu_k \\ \sigma_k^2 &= \frac{1}{\sum_{i,j} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk})} \sum_{i,j} E_{\bar{\theta}}(\rho_{ik}\kappa_{jk}) (y_{ij} - \mu_k - \alpha_{ik} - \beta_{jk})^2 \\ \pi_{cb(k)} &\propto \sum_{i,j} p(\rho_{ik}\kappa_{jk} = 1 | y_{ij}, \bar{\theta}).\end{aligned}$$

C. The penalized plaid model : the full conditionals

The labels

Note that the likelihood may be written as

$$\begin{aligned}\exp \left\{ -\frac{1}{2} \sum_{i,j} (1 - \gamma_{ij}) \left(\left\{ \frac{y_{ij} - \sum_k \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk})}{\sigma(\rho_i, \kappa_j)} \right\}^2 + \log \sigma(\rho_i, \kappa_j)^2 \right) \right. \\ \left. - \frac{1}{2} \sum_{i,j} \gamma_{ij} \left(\left(\frac{y_{ij} - \mu_0}{\sigma_0} \right)^2 + \log \sigma_0^2 \right) \right\}.\end{aligned}$$

Let the bicluster k , be fixed. Define the variables $z_{ijk} = y_{ij} - \sum_{k' \neq k} \rho_{ik'}\kappa_{jk'}(\mu_{k'} - \alpha_{ik'} - \beta_{jk'})$, $\alpha_k = (\alpha_{ik})_{i \in I_k} \in \mathbb{R}^{r_k}$, $\beta_k = (\beta_{jk})_{j \in J_k} \in \mathbb{R}^{c_k}$.

To find the full conditional of the labels, say ρ_{ik} , we use the fact that

$$\begin{aligned}y_{ij} - \sum_{k'=1}^K \rho_{ik'}\kappa_{jk'}(\mu_{k'} + \alpha_{ik'} + \beta_{jk'}) &= z_{ijk} - \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk}) \\ &= \rho_{ik}\kappa_{jk}(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk}) + (1 - \rho_{ik})z_{ijk} + \rho_{ik}(1 - \kappa_{jk})z_{ijk}.\end{aligned}$$

Note that $\gamma_{ij} = \prod_{k=1}^K (1 - \rho_{ik}\kappa_{jk})$. For a given k , we will write $\gamma_{ijk} = \prod_{\substack{k'=1 \\ k' \neq k}}^K (1 -$

$\rho_{ik'}\kappa_{jk'}$). Then $\rho_{ik}\kappa_{jk}(1 - \gamma_{ij}) = \rho_{ik}\kappa_{jk} - \gamma_{ijk}(1 - \rho_{ik}\kappa_{jk})\rho_{ik}\kappa_{jk} = \rho_{ik}\kappa_{jk}$. We have

$$\begin{aligned} & \sum_j (1 - \gamma_{ij}) \frac{(z_{ijk} - \rho_{ik}\kappa_{jk}(\mu_k + \alpha_{ik} + \beta_{jk}))^2}{\sigma^2(\rho_i * \kappa_j)} \\ &= \rho_{ik} \sum_{j \in J_k} \frac{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma^2(\rho_i * \kappa_j)} \\ & \quad + (1 - \rho_{ik}) \sum_j (1 - \gamma_{ij}) \frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)} + \rho_{ik} \sum_{j \notin J_k} (1 - \gamma_{ij}) \frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)} \\ &= \rho_{ik} \sum_{j \in J_k} \frac{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma^2(\rho_i * \kappa_j)} \\ & \quad + (1 - \rho_{ik}) \sum_{j \in J_k} (1 - \gamma_{ijk}) \frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)} + \sum_{j \notin J_k} (1 - \gamma_{ijk}) \frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)}. \end{aligned}$$

As before, let θ denote the set of parameters of the model. Define

$$\begin{aligned} A_{ik} &= \exp \left\{ -\frac{1}{2} \sum_{j \in J_k} \frac{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2}{\sigma^2(\rho_i * \kappa_j)} \right\} \left(\prod_{j \in J_k} \sigma^2(\rho_i * \kappa_j) \right)^{-1/2}, \\ B_{ik} &= \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{j \in J_k} \gamma_{ijk}(y_{ij} - \mu_0)^2 \right\} \left(\sigma_0^2 \right)^{-\sum_{j \in J_k} \gamma_{ijk}/2}, \\ C_{ik} &= \exp \left\{ -\frac{1}{2} \sum_{j \in J_k} (1 - \gamma_{ijk}) \left(\frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)} + \log \sigma^2(\rho_i * \kappa_j) \right) \right\}, \\ D_{ik, \rho_{ik}} &= \exp \left\{ \frac{1}{2} \sum_{j \notin J_k} (1 - \gamma_{ijk}) \left(\frac{z_{ijk}^2}{\sigma^2(\rho_i * \kappa_j)} + \log \sigma^2(\rho_i * \kappa_j) \right) \right\}. \end{aligned}$$

Also let $\rho_{(-ik)}$ denote the set of all row labels except ρ_{ik} . From the above equation it is straightforward to verify that the full conditionals of ρ_{ik} satisfy

$$p(\rho_{ik} | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta) \propto A_{ik}^{\rho_{ik}} B_{ik}^{1-\rho_{ik}} C_{ik}^{1-\rho_{ik}} D_{ik, \rho_{ik}} \pi(\rho_{ik}),$$

where $\pi(\rho_{ik}) = \exp\{-\lambda \sum_j (\sum_{\substack{k'=1 \\ k' \neq k}}^K \rho_{ik'}\kappa_{jk'} + \gamma_{ijk} + (1 - \gamma_{ijk})\kappa_{jk}\rho_{ik} - 1)\}$. In particular, the ratio $p(\rho_{ik} = 1 | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta) / p(\rho_{ik} = 0 | \{y_{ij}\}, (\rho_{(-ik)}, \kappa), \theta)$ is given by

$$A_{ik} B_{ik}^{-1} C_{ik}^{-1} D_{ik,1} D_{ik,0}^{-1} \exp\{-\lambda \sum_j (1 - \gamma_{ijk})\kappa_{jk}\}.$$

The term $D_{ik, \rho_{ik}}$ may be ignored for models whose variances do not depend on (i, j) . In particular, for the plaid model, the logarithm of this ratio is $-\frac{1}{2\sigma^2} \sum_{j \in J_k} \{(z_{ijk} - \mu_k - \alpha_{ik} - \beta_{jk})^2 + (1 - \gamma_{ijk})z_{ijk}^2 + \gamma_{ijk}(y_{ij} - \mu_0)^2\} - \lambda \sum_j (1 - \gamma_{ijk})\kappa_{jk}$. The full conditional for κ_{jk} 's are found in a similar way by symmetry.

The row and column effects

Define the matrices $R_k = \text{diag}(\sum_{j \in J_k} \sigma^{-2}(\rho_i, \kappa_j))$, and $C_k = \text{diag}(\sum_{i \in I_k} \sigma^{-2}(\rho_i, \kappa_j))$. Let $\mathbf{1}_m$ denote the vector of all 1's in $\mathbb{R}^m$. Since the variance of α_k is given by $\sigma_\alpha^2 V_k = \sigma_\alpha^2 (I_{r_k} - \frac{1}{r_k} \mathbf{1}_{r_k} \mathbf{1}_{r_k}')$, we may write $\alpha_k = V_k a_k$ for a random vector $a_k \sim N(0, \sigma_\alpha^2 I_{r_k})$. It is easy to verify that the full conditional of a_k is a multivariate normal with mean $\mu_{a,k}$ and variance $\Sigma_{a,k}$ given by

$$\begin{aligned}\mu_{a,k} &= (V_k R_k V_k + \sigma_\alpha^{-2} I_{r_k})^{-1} V_k z_{\alpha,k}, \\ \Sigma_{a,k} &= (V_k R_k V_k + \sigma_\alpha^{-2} I_{r_k})^{-1},\end{aligned}$$

where $z_{\alpha,k} = (\sum_{j \in J_k} (z_{ij} - \mu_k - \beta_{jk}) / \sigma^2(\rho_i, \kappa_j))_{i \in I_k}$. Similarly, let $U_k = (I_{c_k} - \frac{1}{c_k} \mathbf{1}_{c_k} \mathbf{1}_{c_k}')$. We may write $\beta_k = U_k b_k$ for a random vector $b_k \sim N(0, \sigma_\beta^2 I_{c_k})$. It is easy to verify that the full conditional of b_k is a multivariate normal with mean $\mu_{b,k}$ and variance $\Sigma_{b,k}$ given by

$$\begin{aligned}\mu_{b,k} &= (U_k C_k U_k + \sigma_\beta^{-2} I_{c_k})^{-1} U_k z_{\beta,k}, \\ \Sigma_{b,k} &= (U_k C_k U_k + \sigma_\beta^{-2} I_{c_k})^{-1},\end{aligned}$$

where $z_{\beta,k} = (\sum_{i \in I_k} (z_{ij} - \mu_k - \alpha_{ik}) / \sigma^2(\rho_i, \kappa_j))_{j \in J_k}$.

For the plaid model $\sigma(\rho_i, \kappa_j) = \sigma$, and for the model of Cheng and Church [1], $\sigma(\rho_i, \kappa_j) = \sigma_k$. In both cases the variance is constant on each bicluster. Therefore, for these models, $R_k = \sigma_k^{-2} c_k I_{r_k}$ and $C_k = \sigma_k^{-2} r_k I_{c_k}$. Hence the conditional means and variances for a_k and b_k become

$$\begin{aligned}\mu_{a,k} &= \left(\frac{c_k}{\sigma_k^2} + \frac{1}{\sigma_\alpha^2} \right)^{-1} \frac{c_k}{\sigma_k^2} (\bar{z}_{i \cdot k} - \bar{z}_k)_{i \in I_k}, \\ \Sigma_{a,k} &= \left(\frac{c_k}{\sigma_k^2} + \frac{1}{\sigma_\alpha^2} \right)^{-1} \left(I_{r_k} + \frac{c_k}{r_k} \frac{\sigma_\alpha^2}{\sigma_k^2} \mathbf{1}_{r_k} \mathbf{1}_{r_k}' \right), \\ \mu_{b,k} &= \left(\frac{r_k}{\sigma_k^2} + \frac{1}{\sigma_\beta^2} \right)^{-1} \frac{r_k}{\sigma_k^2} (\bar{z}_{\cdot j k} - \bar{z}_k)_{j \in J_k}, \\ \Sigma_{b,k} &= \left(\frac{r_k}{\sigma_k^2} + \frac{1}{\sigma_\beta^2} \right)^{-1} \left(I_{c_k} + \frac{r_k}{c_k} \frac{\sigma_\beta^2}{\sigma_k^2} \mathbf{1}_{c_k} \mathbf{1}_{c_k}' \right),\end{aligned}$$

where $\bar{z}_k$ denotes the mean of the values of z_{ijk} in the bicluster k , and $\bar{z}_{i \cdot k} = \sum_{j \in J_k} z_{ijk} / c_k$, and $\bar{z}_{\cdot j k} = \sum_{i \in I_k} z_{ijk} / r_k$.

It can be easily shown that the full conditionals of the means μ_k , $k = 0, 1, \dots, K$ are also normal distributions with means and variances given by

$$\begin{aligned}\mu_{\mu,k} &= \left(\frac{1}{\sigma_\mu^2} + \sum_{(i,j) \in B_k} \frac{1}{\sigma^2(\rho_i, \kappa_j)} \right)^{-1} \sum_{(i,j) \in B_k} \frac{z_{ij} - \alpha_{ik} - \beta_{jk}}{\sigma^2(\rho_i, \kappa_j)}, \\ \Sigma_{\mu,k} &= \left(\frac{1}{\sigma_\mu^2} + \sum_{(i,j) \in B_k} \frac{1}{\sigma^2(\rho_i, \kappa_j)} \right)^{-1}.\end{aligned}$$

Again, for the plaid and Cheng and Church models, the means and variances

simplify to

$$\mu_{\mu,k} = \left(\frac{1}{\sigma_\mu^2} + \frac{n_k}{\sigma_k^2} \right)^{-1} \frac{n_k}{\sigma_k^2} \bar{z}_k, \quad \Sigma_{\mu,k} = \left(\frac{1}{\sigma_\mu^2} + \frac{n_k}{\sigma_k^2} \right)^{-1}.$$

Note that when σ_μ^2 , σ_α^2 , and σ_β^2 tend to infinity we obtain the hard-EM (or ICM) estimators.

The variances

Let $z_{ij} = y_{ij} - \sum_{k'} \rho_{ik'} \kappa_{jk'} (\mu_{k'} + \alpha_{ik'} + \beta_{jk'})$. The full conditionals of the variances σ_k^2 are proportional to

$$\exp \left\{ -\frac{1}{2} \left(\sum_{(i,j) \in B_k} \frac{z_{ij}^2 + \nu s^2}{\sigma_k^2} + \frac{\nu s^2}{\sigma_k^2} \right) - \frac{1}{2} n_k \log(\sigma_k^2) - \left(\frac{\nu}{2} + 1 \right) \log \sigma_k^2 \right\}.$$

If we suppose that there is no overlapping among the biclusters, we obtain the Cheng and Church model [1]. The corresponding full conditional of σ_k^2 is an inverse- χ^2 distribution with scale $(\nu s^2 + \sum_{(i,j) \in B_k} z_{ij}^2)/(\nu + n_k)$, and $\nu + n_k$ degrees of freedom. If instead we suppose that $\sigma(\rho_i, \kappa_j) = \sigma$ independently of the cell (i, j) (i.e., $\sigma_k = \sigma$ for all $k = 0, 1, \dots, K$), then we obtain the full conditional of σ^2 for the plaid model. This is also an inverse- χ^2 distribution, but this time with scale $(\nu s^2 + \sum_{i,j} z_{ij}^2)/(\nu + pq)$, and $\nu + pq$ degrees of freedom.

D. The penalized plaid model: initial values

Finding the initial membership labels (ρ, κ) is a difficult task. Several procedures have been suggested in the literature. We have adopted a technique similar to that of Turner et al. [3]. We run two independent k -means algorithms [4] with $k = 2$: once for the rows and once for the columns. Using the Cartesian product of the resulting k -means row and column labels, we divide the data matrix into four groups. A single initial bicluster is chosen among these four groups according to a variance criterion explained a few lines below. The procedure is repeated as many times as the number of initial biclusters needed. A single initial bicluster is chosen after each application of the independent row and column k -means algorithms. The elements of the biclusters already chosen are *masked* by replacing their original values y_{ij} by random values. This is done so that in the next iteration a different group may be chosen. The masking procedure is not new. It has been used before by Sheng, Moreau and De Moor [2] to determinate multiple biclusters. The criterion to choose an initial bicluster among the four groups yielded by each iteration is the following. Suppose that the cells of each group follow a random effect additive ANOVA model. That is, on each group g , $y_{ij} = \mu_g + \alpha_{ig} + \beta_{jg} + \epsilon_{ij}$, $g = 1, 2, 3, 4$. The standard moment estimates of the variances are

$$\hat{\sigma}_{g\alpha}^2 = \frac{MSS_g(\alpha) - MSS_g(e)}{c_g}, \quad \hat{\sigma}_{g\beta}^2 = \frac{MSS_g(\beta) - MSS_g(e)}{r_g}, \quad \hat{\sigma}_{ge}^2 = MSS_g(e), \quad (1)$$

where r_g is the number of rows in the g -th group, c_g is the number of columns, and $MSS_g(e)$, $MSS_g(\alpha)$ and $MSS_g(\beta)$ are the corresponding mean sum of squares for error, rows and columns, respectively. We select as an initial bicluster the group g

that maximizes $(\hat{\sigma}_{g\alpha}^2 + \hat{\sigma}_{g\beta}^2)/\hat{\sigma}_{ge}^2$. For each initial bicluster, the parameters μ_g , α_{ig} and β_{jg} are initialized as $\bar{y}_{..}$, $\bar{y}_{i.} - \bar{y}_{..}$, $\bar{y}_{.j} - \bar{y}_{..}$, respectively, where $y_{..}$, $\bar{y}_{i.}$, $\bar{y}_{.j}$ stand for the overall bicluster mean, the bicluster i -th row mean, and the bicluster j -th column mean, respectively. The parameter μ_0 is estimated as the arithmetic mean of the zero-bicluster. The variance σ^2 is initialized as the mean sum of squares of all the residuals.

E. Biclustering algorithm packages

The following table gives the URLs associated with the biclustering packages used in the comparison study presented in Section 5 of the main body of this paper. The package names are given in parenthesis under the associated algorithm names.

Algorithm	Description and Address
FABIA (fabia)	R package: Factor Analysis for Bicluster Acquisition www.bioconductor.org/packages/release/bioc/html/fabia.html
SAMBA (EXPANDER)	EXpression Analyzer and DisplayER: java-based tool for analysis of gene expression and NGS ¹ data acgt.cs.tau.ac.il/expander/overview.html
TUR (biclust)	R package: the plaid model implementation of Turner and al. cran.r-project.org/web/packages/biclust/
CC (biclust)	R package: original method suggested by Cheng and Church to fit a mixture model cran.r-project.org/web/packages/biclust/
Spectral (biclust)	R package: uses a singular value decomposition cran.r-project.org/web/packages/biclust/
Penalized plaid (penalizedplaid)	java package: The models are: (1) the non-overlapping bicluster model; (2) the plaid model; and (3) the penalized plaid model. The estimation methods are: (a) the Gibbs sampler and (b) the Metropolis-Hastings algorithm www.dms.umontreal.ca/~murua/

References

- [1] Y. Cheng and G. Church, *Biclustering of expression data*, Int. Conf. Intelligent Systems for Molecular Biology 12 (2000), pp. 61–86.

¹Next Generation Sequencing

- [2] Q. Sheng, Y. Moreau, and B. De Moor, *Biclustering microarray data by Gibbs sampling*, Bioinformatics 19 (2003), pp. ii196–ii205.
- [3] H. Turner, T. Bailey, and W. Krzanowski, *Improved biclustering of microarray data demonstrated through systematic performance tests*, Computational Statistics & Data Analysis 48 (2005), pp. 235–254.
- [4] J.H. Ward, *Hierarchical groupings to optimize an objective function*, J. American Statistical Association 58 (1963), pp. 234–244.