GIBBS SAMPLING

The posterior distribution for which we obtain samples via the Gibbs sampling procedure can be written as:

$p(\mathbf{g}, \mathbf{c}, \boldsymbol{\theta} \mid \mathbf{X}, \boldsymbol{\psi})$

Here, $\mathbf{g}$ is the vector of binary gene labels, $\mathbf{c}$ the vector of binary condition labels, $\boldsymbol{\theta}$ the vector of model parameters, $\boldsymbol{\psi}$ the total set of hyperparameters (κ, ν, s, φ, ξ_{g0}, ξ_{g1}, ξ_{c0}, ξ_{c1}), $\mathbf{X}$ the expression data matrix (with rows referring to genes, and columns to conditions), $B(\xi_{g1}, \xi_{g0})$ the Beta prior distribution on the probability that a gene belongs to the bicluster, $B(\xi_{c1}, \xi_{c0})$ the Beta prior distribution on the probability that a condition belongs to the bicluster,

Derivation of full conditional Bernouilli distribution for gene and condition labels

1. Assume that the prior probability that a gene (condition) label equals one is given by a Beta distribution:

   $B(\xi_{g1}, \xi_{g0})$ and $B(\xi_{c1}, \xi_{c0})$

2. The probability distribution of a collection $\mathbf{g}$ of n binary gene labels is then found by integrating out over this (conjugate) Beta prior distribution:

$$
\begin{aligned}
& p(\mathbf{g} \mid \xi_{g1}, \xi_{g0}) \\
&= \int p(\mathbf{g} \mid x)\, B(x, \xi_{g1}, \xi_{g0})\, dx \\
&= \int \prod_{i=1}^{n} Bern(g_i, x)\, B(x, \xi_{g1}, \xi_{g0})\, dx \\
&= \int x^{\|\mathbf{g}\|_1} (1-x)^{n-\|\mathbf{g}\|_1}\, B(x, \xi_{g1}, \xi_{g0})\, dx \\
&= \frac{\Gamma(\xi_{g0} + n - \|\mathbf{g}\|_1)\, \Gamma(\xi_{g1} + \|\mathbf{g}\|_1)}{\Gamma(\xi_{g0} + \xi_{g1} + n)}
\end{aligned}
$$

   n refers to the total number of genes and $\|\mathbf{g}\|_1$ to the one norm of the current (binary) gene label vector (number of genes currently in the bicluster).

3. The derivation for the length m vector of binary condition labels $\mathbf{c}$ is analogous, and leads to:

$$
p(\mathbf{c} \mid \xi_{c1}, \xi_{c0}) = \frac{\Gamma(\xi_{c0} + m - \|\mathbf{c}\|_1)\, \Gamma(\xi_{c1} + \|\mathbf{c}\|_1)}{\Gamma(\xi_{c0} + \xi_{c1} + m)}
$$

4. Now consider the full conditional distributions of the gene labels. The conditional for gene label *i* is modeled by a Bernouilli distribution with parameter α_i, that gives the probability that the label equals one:

$$\begin{aligned}\alpha_i &= p(g_i = 1 \mid \mathbf{X}, \mathbf{g}_{\neq i}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(g_i = 1 \mid \mathbf{X}, \mathbf{g}_{\neq i}, \mathbf{c}, \boldsymbol{\theta}, \xi_{g1}, \xi_{g0}) \\ &= \frac{p(\mathbf{X}, g_i = 1, \mathbf{g}_{\neq i} \mid \mathbf{c}, \boldsymbol{\theta}, \xi_{g1}, \xi_{g0})}{p(\mathbf{X}, \mathbf{g}_{\neq i} \mid \mathbf{c}, \boldsymbol{\theta}, \xi_{g1}, \xi_{g0})} = \frac{p(\mathbf{X}, g_i = 1, \mathbf{g}_{\neq i} \mid \mathbf{c}, \boldsymbol{\theta}, \xi_{g1}, \xi_{g0})}{N} \\ &= \frac{1}{N} p(g_i = 1, \mathbf{g}_{\neq i} \mid \xi_{g1}, \xi_{g0}) p(\mathbf{X}_{\mathbf{i}\cdot} \mid g_i = 1, \mathbf{c}, \boldsymbol{\theta}) \prod_{\substack{k=1 \\ k \neq i}}^{n} p(\mathbf{X}_{\mathbf{k}\cdot} \mid g_k, \mathbf{c}, \boldsymbol{\theta}) \\ &= \frac{1}{N} p(g_i = 1, \mathbf{g}_{\neq i} \mid \xi_{g1}, \xi_{g0}) \prod_{j \mid c_j = 1} p^{\text{bcl}}(x_{ij}) \prod_{j \mid c_j = 0} p^{\text{bgd},0}(x_{ij}) \prod_{\substack{k=1 \\ k \neq i}}^{n} p(\mathbf{X}_{\mathbf{k}\cdot} \mid g_k, \mathbf{c}, \boldsymbol{\theta})\end{aligned}$$

In the above, $\mathbf{X}_{\mathbf{k}\cdot}$ indicates the *k*-th row in the expression matrix $\mathbf{X}$.

5. Similarly, an expression for $1-\alpha_i$ can be derived:

$$\begin{aligned}1 - \alpha_i &= p(g_i = 0 \mid \mathbf{X}, \mathbf{g}_{\neq i}, \mathbf{c}, \boldsymbol{\theta}, \boldsymbol{\psi}) \\ &= \frac{1}{N} p(g_i = 0, \mathbf{g}_{\neq i} \mid \xi_{g1}, \xi_{g0}) \prod_{j \mid c_j = 1} p^{\text{bgd},1}(x_{ij}) \prod_{j \mid c_j = 0} p^{\text{bgd},0}(x_{ij}) \prod_{\substack{k=1 \\ k \neq i}}^{n} p(\mathbf{X}_{\mathbf{k}\cdot} \mid g_k, \mathbf{c}, \boldsymbol{\theta})\end{aligned}$$

6. Now, the ratio of those two expressions has a simple form and an elegant interpretation, as discussed in some detail in the main text:

$$\begin{aligned}\frac{\alpha_i}{1-\alpha_i} &= \frac{p(g_i = 1, \mathbf{g}_{\neq i} \mid \xi_{g1}, \xi_{g0})}{p(g_i = 0, \mathbf{g}_{\neq i} \mid \xi_{g1}, \xi_{g0})} \frac{\prod_{j \mid c_j = 1} p^{\text{bcl}}(x_{ij})}{\prod_{j \mid c_j = 1} p^{\text{bgd},1}(x_{ij})} \\ &= \frac{\xi_{g1} + \left\| \mathbf{g}_{\neq i} \right\|_1}{\xi_{g0} + n - \left\| \mathbf{g}_{\neq i} \right\|_1} \prod_{j \mid c_j = 1} \frac{p^{\text{bcl}}(x_{ij})}{p^{\text{bgd},1}(x_{ij})}\end{aligned}$$

The last equation is a direct consequence of the result described under 2.

7. The odds for the Bernouilli parameters of the condition label conditionals can be obtained similarly:

$$\beta_j = p(c_j = 1 \mid \mathbf{X}, \mathbf{c}_{\neq j}, \mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\psi}) = p(c_j = 1 \mid \mathbf{X}, \mathbf{c}_{\neq i}, \mathbf{g}, \boldsymbol{\theta}, \xi_{c1}, \xi_{c0})$$

$$= \frac{p(\mathbf{X}, c_j = 1, \mathbf{c}_{\neq i} \mid \mathbf{g}, \boldsymbol{\theta}, \xi_{c1}, \xi_{c0})}{p(\mathbf{X}, \mathbf{c}_{\neq j} \mid \mathbf{g}, \boldsymbol{\theta}, \xi_{c1}, \xi_{c0})} = \frac{p(\mathbf{X}, c_j = 1, \mathbf{c}_{\neq j} \mid \mathbf{g}, \boldsymbol{\theta}, \xi_{c1}, \xi_{c0})}{M}$$

$$= \frac{1}{M} p(c_j = 1, \mathbf{c}_{\neq j} \mid \xi_{c1}, \xi_{c0}) p(\mathbf{X}_{\mathbf{.j}} \mid c_j = 1, \mathbf{g}, \boldsymbol{\theta}) \prod_{\substack{k=1 \\ k \neq i}}^{m} p(\mathbf{X}_{\mathbf{.k}} \mid c_k, \mathbf{g}, \boldsymbol{\theta})$$

$$= \frac{1}{M} p(c_j = 1, \mathbf{c}_{\neq j} \mid \xi_{c1}, \xi_{c0}) \prod_{i \mid g_i = 1} p^{\text{bcl}}(x_{ij}) \prod_{i \mid g_i = 0} p^{\text{bgd},1}(x_{ij}) \prod_{\substack{k=1 \\ k \neq i}}^{m} p(\mathbf{X}_{\mathbf{.k}} \mid c_k, \mathbf{g}, \boldsymbol{\theta})$$

$$1 - \beta_j = p(c_j = 0 \mid \mathbf{X}, \mathbf{c}_{\neq j}, \mathbf{g}, \boldsymbol{\theta}, \boldsymbol{\psi}, \xi_{g1}, \xi_{g0}, \xi_{c1}, \xi_{c0})$$

$$= \frac{1}{M} p(c_j = 0, \mathbf{c}_{\neq j} \mid \xi_{c1}, \xi_{c0}) \prod_{i \mid g_i = 1} p^{\text{bgd},0}(x_{ij}) \prod_{i \mid g_i = 0} p^{\text{bgd},0}(x_{ij}) \prod_{\substack{k=1 \\ k \neq i}}^{m} p(\mathbf{X}_{\mathbf{.k}} \mid c_k, \mathbf{g}, \boldsymbol{\theta})$$

$$\frac{\beta_j}{1 - \beta_j} = \frac{p(c_j = 1, \mathbf{c}_{\neq j} \mid \xi_{c1}, \xi_{c0})}{p(c_j = 0, \mathbf{c}_{\neq j} \mid \xi_{c1}, \xi_{c0})} \frac{\prod_{i \mid g_i = 1} p^{\text{bcl}}(x_{ij}) \prod_{i \mid g_i = 0} p^{\text{bgd},1}(x_{ij})}{\prod_{i=1}^{n} p^{\text{bgd},0}(x_{ij})}$$

$$= \frac{\xi_{c1} + \left\| \mathbf{c}_{\neq j} \right\|_1}{\xi_{c0} + m - \left\| \mathbf{c}_{\neq j} \right\|_1} \prod_{i \mid g_i = 1} \frac{p^{\text{bcl}}(x_{ij})}{p^{\text{bgd},0}(x_{ij})} \prod_{i \mid g_i = 0} \frac{p^{\text{bgd},1}(x_{ij})}{p^{\text{bgd},0}(x_{ij})}$$

The extra factor is a result of the asymmetrical nature of the patterns we are looking for ('striped' biclusters rather than constant-value biclusters).

Derivation of full conditional Bernouilli distribution for model parameters

Due to the conjugacy, the full conditional distributions for the model are in the same form as the priors.

1. One can easily derive that for the mean, the product of the normal likelihood with the normal prior (conditioned on σ) leads to a posterior normal distribution for the mean.

2. Similarly, combined with a normal likelihood, the Inverse χ^2 prior on σ^2 gives rise to an Inverse χ^2 posterior on σ^2.

This interpretation that directly motivates the parameterization of the prior, as it allows to interpret both κ and ν as a number of prior observations (sometimes referred to as pseudocounts) associated with the prior. For a derivation of the corresponding formulas,

we refer the reader to standard textbooks on Bayesian statistics (Gelman A.B. et al., 2004).

Note on the use of two background models

Strictly speaking there are two background models rather than one single model, as shown in figure REF. If condition label *j* is zero, the entire *j* th column is captured by one background model (bgd,0). On the other hand, if the *j* th column label equals one, the background model that describes the expression of those background genes should only be constructed with the background genes (bgd,1). The difference between those models only shows up in the derivations for the full conditional distributions of the condition labels (see above).

In most cases this distinction is primarily a technical issue and the difference between bgd,0 and bgd,1 only matters when very large biclusters occur.

SCORING MEASURES

Biclustering results were scored with *module recovery* and *bicluster relevance* scores as described in (Prelic et al., 2006). The *module recovery* score indicates how well the gene content of the 'ideal' modules is on average reflected in the (best matching bicluster of the) bicluster results:

$$S_{MR}(M_1, M_2) = \frac{1}{M_1} \sum_{(G_1,C_1)\in M_1} \max_{(G_2,C_2)\in M_2} \frac{|G_1 \cap G_2|}{|G_1 \cup G_2|}$$

In the above, M_1 indicates the set of biclusters or modules that are considered 'ideal' or 'real', M_2 indicates the set of biclusters that are really returned by the algorithm. Hence, module recovery runs over all 'ideal' biclusters and averages out the match scores (in terms of gene content G) of the best matching biclusters from the obtained biclustering results.

The *bicluster relevance* score S_{BR} is obtained by interchanging the roles of the 'ideal' and reported biclusters. It is related to the relevance of the set of modules in the output. In other words, it penalizes 'incorrect' modules in the output.

REFERENCES

Gelman A.B. et al. (2004) *Bayesian data analysis*. Chapman & Hall/CRC.

Prelic,A. et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics.*, **22**, 1122-1129.