1 The GenMiR++ model: learning and inference

GenMiR++ models the relationship between the miRNA expression data $\mathbf{Z}$ and the gene expression data $\mathbf{X}$ given a set of candidate miRNA-mRNA target interactions $\mathbf{C}$ using a probabilistic generative model that depends on nuisance variables $\{\mathbf{\Lambda}, \mathbf{\Gamma}\}$ and parameters $\Theta = \{\boldsymbol{\mu}, \mathbf{\Sigma}, \pi, \alpha\}$ and δ^2. The variable $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_K\}$ is a vector of positive regulatory weights that represent the efficiency of down-regulation by miRNAs 1 through K. The variable $\mathbf{\Gamma} = \text{diag}(\gamma_1, \cdots, \gamma_T)$ is a $T \times T$ diagonal matrix of positive tissue-scaling parameters that account for normalization differences in the expression data for tissues 1 through T. These nuisance variables are integrated out during learning and inference in the GenMiR++ model. We do a point estimation of the parameters $\Theta = \{\boldsymbol{\mu}, \mathbf{\Sigma}, \pi, \alpha\}$ and set $\delta^2 = 0.01$.

Having defined the above parameters and variables, we can write the probabilities in our model, conditioned on the expression of miRNAs $\mathbf{X}$ and a set of candidate miRNA targets $\mathbf{C}$, as

$$p(\mathbf{x}_g|\mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \Theta) = \mathcal{N}(\mathbf{x}_g; \boldsymbol{\mu} - \sum_k \lambda_k s_{gk} \mathbf{\Gamma} \mathbf{z}_k, \mathbf{\Sigma}) \tag{1}$$

$$p(\mathbf{S}|\mathbf{C}, \Theta) = \prod_{(g,k)} p(s_{gk}|\mathbf{C}, \Theta) = \prod_{(g,k)|c_{gk}=0} (1 - s_{gk}) \prod_{(g,k)|c_{gk}=1} \pi^{s_{gk}} (1 - \pi)^{1 - s_{gk}} \tag{2}$$

$$p(\mathbf{\Gamma}|\Theta) = \prod_{t=1}^{T} p(\gamma_t|\Theta) = \prod_{t=1}^{T} \mathcal{N}(\gamma_t; 1, \delta^2) \tag{3}$$

$$p(\mathbf{\Lambda}|\Theta) = \prod_{k=1}^{K} p(\lambda_k|\Theta) = \prod_{k=1}^{K} \frac{1}{\alpha} \exp\Big(- \frac{\lambda_k}{\alpha} \Big) \tag{4}$$

$$p(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{Z}, \Theta) = p(\mathbf{S}|\mathbf{C}, \Theta) p(\mathbf{\Gamma}|\Theta) p(\mathbf{\Lambda}|\Theta) \prod_g p(\mathbf{x}_g|\mathbf{Z}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}, \Theta) \tag{5}$$

$$p(\mathbf{X}|\mathbf{C}, \mathbf{Z}, \Theta) = \sum_{\mathbf{S}} \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} p(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{Z}, \Theta) \, d\mathbf{\Lambda} \, d\mathbf{\Gamma} \tag{6}$$

where $\mathcal{N}(x; m, v)$ is the density of x under a normal distribution with mean x and covariance matrix v. If x is a scalar then v is simply a scalar variance.

Having presented the above model for miRNA post-transcriptional regulation, we would now like to infer the settings for the s_{gk} variables by computing the posterior probability $p(\mathbf{S}|\mathbf{X}, \mathbf{Z}, \mathbf{C}, \Theta) \propto \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} p(\mathbf{X}, \mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{Z}, \mathbf{C}, \Theta) \, d\mathbf{\Lambda} \, d\mathbf{\Gamma}$. Note that to compute this posterior, we have to integrate out the nuisance variables $\mathbf{\Lambda}, \mathbf{\Gamma}$ and sum over an exponential number of configurations of the

miRNA-mRNA targetting variables $\mathbf{S}$ as such computing this posterior exactly is intractable. We instead use a variational Bayesian technique to both approximate the posterior over $\mathbf{S}$ and find a setting of Θ which maximises a lower bound on the likelihood of $\mathbf{X}$ given $\mathbf{Z}$ and $\mathbf{C}$.

Variational inference methods approximate probabilistic inference by approximating the posterior distribution $p(u|v, \Theta)$ over the unobserved variables u given the observed variables v (and parameters Θ) with a surrogate distribution $q(u|\mathbf{\Psi})$ that depends on variational parameters $\mathbf{\Psi}$. This surrogate distribution is chosen to have an analytically tractable form so that when it is substituted for the true posterior, any calculations that depend on the posterior, *e.g.* parameter updates for Θ, become tractable.

Variational inference proceeds by fitting $q(u, \mathbf{\Psi})$ to $p(u|v, \Theta)$ by optimizing a quantity called the *variational free energy* [Neal and Hinton 1998] with respect to $\mathbf{\Psi}$. This free energy is equal to the log likelihood of $\mathbf{v}$, $\log\ p(\mathbf{v} \mid \Theta)$, minus the Kullback-Leibler (KL) divergence (or relative entropy) between q and p. So maximizing the free energy is equivalent to minimizing the KL-divergence between q and p. However, the former task can often be made tractable by a careful choice of the functional form of q, while the latter task cannot.

In addition to calculating an approximation to the posterior, we would also like to find a setting of our model parameters Θ that maximizes the likelihood of the data. The free energy can also be used for this task. In particular, for a fixed q specified by $\mathbf{\Psi}$, maximizing the free energy with respect to Θ is equivalent to maximizing the log likelihood of the data. In order to jointly maximize the free energy with respect to Θ and $\mathbf{\Psi}$, we iterate between maximizing each individually. This procedure is often called the Expectation Maximization algorithm. Because we integrate over distributions of the latent nuisance variables $\mathbf{\Gamma}$ and $\mathbf{\Lambda}$ when maximizing the log likelihood instead of maximizing them along with our parameters, we are said to be using a variational Bayesian learning procedure [Attias 1999]. More details on variational inference and Bayesian learning are available from [Bishop 2006].

In the GenMiR++ model, $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{C}$ are observed; $\mathbf{S}$, $\mathbf{\Lambda}$ and $\mathbf{\Gamma}$ are the unobserved variables whose posterior distribution we are approximating; and Θ is the parameter set. We can write the variational free energy, $F(\mathbf{\Theta}, \mathbf{\Psi})$ as

$$F(\mathbf{\Theta}, \mathbf{\Psi}) = \sum_{\mathbf{S}} \int_{\mathbf{\Gamma}} \int_{\mathbf{\Lambda}} q(\mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{C}, (\Psi)) \log \frac{q(\mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{C}, (\Psi))}{p(\mathbf{X}, \mathbf{S}, \mathbf{\Gamma}, \mathbf{\Lambda}|\mathbf{C}, \mathbf{Z}, \Theta)}\, d\mathbf{\Lambda}\, d\mathbf{\Gamma}, \quad (7)$$

where $\mathbf{\Psi} = \{\boldsymbol{\beta}, \boldsymbol{\nu}, \mathbf{\Omega}, \mathbf{\Phi}\}$.

To allow for tractable learning, we use a q-distribution with a mean-field factorization, *i.e.*, we define $q(\mathbf{S}, \mathbf{\Lambda}, \mathbf{\Gamma}|\mathbf{C}, \mathbf{\Psi}) = q(\mathbf{S}|\mathbf{C}, \boldsymbol{\beta}) q(\mathbf{\Lambda}|\boldsymbol{\nu}) q(\mathbf{\Gamma}|\mathbf{\Omega}, \mathbf{\Phi})$ where

$$
\begin{aligned}
q(\mathbf{S}|\mathbf{C},\boldsymbol{\beta}) &= \prod_{g,k} q(s_{gk}|c_{gk},\beta_{gk}) = \prod_{(g,k)|c_{gk=1}} \beta_{gk}^{s_{gk}}(1-\beta_{gk})^{1-s_{gk}} \\
q(\boldsymbol{\Lambda}|\boldsymbol{\nu}) &= \prod_{k} q(\lambda_k|\nu_k) = \prod_{k} \frac{1}{\nu_k}\exp\Big(\frac{-\lambda_k}{\nu_k}\Big) \\
q(\boldsymbol{\Gamma}|\boldsymbol{\Omega},\boldsymbol{\Phi}) &= \prod_{t} q(\gamma_t|\omega_t,\phi_t) = \prod_{t} \mathcal{N}(\gamma_t;\omega_t,\phi_t^2)
\end{aligned} \tag{8}
$$

so under our q-distribution, the targeting indicator variables s_{gk}, the regulatory weights λ_k and the tissue-scaling parameters γ_t are all modelled as being independent of one another given the data. Among our set of variational parameters, β_{gk} can be interpreted as the probability that miRNA k targets mRNA g given the data; ν_k the expected values of the regulatory efficiencies; and ω_t, ϕ_t^2 the means and variances, respectively, of the tissue scaling parameters.

To simplify the presentation of our variational inference and learning algorithm, we will rewrite $F(\boldsymbol{\Theta},\boldsymbol{\Psi})$ by first defining the set of expected sufficient statistics $\boldsymbol{\Omega}, \boldsymbol{\Phi}$, $\mathbf{y}_g$, $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$ as

$$
\begin{aligned}
&\boldsymbol{\Omega} = \mathrm{diag}(\omega_1,\cdots,\omega_T), \qquad \boldsymbol{\Phi} = \mathrm{diag}(\phi_1^2,\cdots,\phi_T^2) \\
&\mathbf{y}_g = \sum_{k:c_{gk}=1} \nu_k\beta_{gk}\mathbf{z}_k \\
&\mathbf{U} = \frac{1}{G}\sum_{g}\Big(\mathbf{x}_g - (\boldsymbol{\mu}-\boldsymbol{\Omega}\mathbf{y}_g)\Big)\Big(\mathbf{x}_g - (\boldsymbol{\mu}-\boldsymbol{\Omega}\mathbf{y}_g)\Big)^{\mathsf{T}} \\
&\mathbf{V} = \frac{1}{G}\sum_{g}\sum_{k:c_{gk}=1} \nu_k^2(2\beta_{gk}-\beta_{gk}^2)(\boldsymbol{\Omega}^2+\boldsymbol{\Phi})\mathbf{z}_k\mathbf{z}_k^{\mathsf{T}} \\
&\mathbf{W} = \frac{1}{G}\boldsymbol{\Phi}\sum_{g}\mathbf{y}_g\mathbf{y}_g^{\mathsf{T}}.
\end{aligned} \tag{9}
$$

Given the above definitions, the free energy $F(\boldsymbol{\Theta},\boldsymbol{\Psi})$ can be written compactly as

$$
\begin{aligned}
F(\boldsymbol{\Theta},\boldsymbol{\Psi}) = &\sum_{(g,k)|c_{gk}=1}\Big(\beta_{gk}\log\frac{\beta_{gk}}{\pi} + (1-\beta_{gk})\log\frac{1-\beta_{gk}}{1-\pi}\Big) \\
&+ \frac{G}{2}\log|\boldsymbol{\Sigma}| + \frac{G}{2}\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}(\mathbf{U}+\mathbf{V}+\mathbf{W})\Big) \\
&+ \frac{1}{2}\left[T\log\delta^2 + \frac{1}{\delta^2}\mathrm{tr}\big((\boldsymbol{\Omega}-\mathbf{I})^2+\boldsymbol{\Phi}\big) - \log|\boldsymbol{\Phi}|\right] \\
&+ \sum_{k}\Big(\frac{\nu_k}{\alpha} - \log\frac{\nu_k}{\alpha}\Big) + \mathrm{const.}
\end{aligned} \tag{10}
$$

We will use variational Bayesian learning to simultaneously fit Θ and approximate the posterior distributions over the targeting indicator variables $q(\mathbf{S}|\boldsymbol{\beta})$.

We will do this, as described above, by iteratively maximizing $F(\boldsymbol{\Theta}, \boldsymbol{\Psi})$ with respect to $\boldsymbol{\beta}$ (variational Bayes E-step), and $\boldsymbol{\Omega}, \boldsymbol{\Phi}, \boldsymbol{\mu}$ (variational Bayes M-step) and with respect to $\boldsymbol{\Theta}$ (parameter optimization step) until convergence to a local maximum of $F(\boldsymbol{\Theta}, \boldsymbol{\Psi})$. The updates of our variational Bayes algorithm are:

Variational Bayes E-step:

$$\forall(g,k)|c_{gk}=1,$$
$$\frac{\beta_{gk}}{1-\beta_{gk}} = \frac{\pi}{1-\pi}\exp\left[-\nu_k \mathbf{z}_k^T \boldsymbol{\Sigma}^{-1}\Big(\boldsymbol{\Omega}\big(\mathbf{x}_g-\boldsymbol{\mu}\big)+\nu_k\big(\boldsymbol{\Omega}^2+\boldsymbol{\Phi}\big)\mathbf{z}_k + \frac{1}{2}\sum_{l\neq k:(g,l)|c_{gl}=1}\nu_l\beta_{gl}\big(\boldsymbol{\Omega}^2+\boldsymbol{\Phi}\big)\mathbf{z}_l\Big)\right] \tag{11}$$

Variational Bayes M-step:

$$\boldsymbol{\Phi} = \mathrm{diag}(\delta^2(\mathbf{I}+\delta^2\mathbf{A})^{-1})$$
$$\boldsymbol{\Omega} = \mathrm{diag}\left(\left(\frac{1}{\delta^2}\mathbf{I}+\mathbf{B}\right)^{-1}\right)\mathrm{diag}\left(\left(\frac{1}{\delta^2}\mathbf{I}+\mathbf{A}\right)^{-1}\right),$$
$$\mathbf{A} = \mathrm{diag}\Big(\boldsymbol{\Sigma}^{-1}\sum_{(g,k)|c_{gk}=1}\nu_k^2(2\beta_{gk}-\beta_{gk}^2)\mathbf{z}_k\mathbf{z}_k^T + \sum_g \mathbf{y}_g\mathbf{y}_g\Big), \mathbf{B} = -\mathrm{diag}\Big(\boldsymbol{\Sigma}^{-1}\sum_g \mathbf{y}_g(\mathbf{x}_g-\boldsymbol{\mu})^T\Big)$$
$$\nu_k = \frac{-b+\sqrt{b^2-4ad}}{2a}, \qquad a = 2\sum_g \beta_{gk}\mathbf{z}_k^T\big(\boldsymbol{\Omega}^2+\boldsymbol{\Phi}\big)\boldsymbol{\Sigma}^{-1}\mathbf{z}_k, b = \frac{1}{\alpha}+\sum_g\beta_{gk}(\mathbf{x}_g-\boldsymbol{\mu})^T\boldsymbol{\Omega}\boldsymbol{\Sigma}^{-1}\mathbf{z}_k, d=-1$$
(12)

Regular Parameter Optimization step:

$$\boldsymbol{\mu} = \arg\max_{\boldsymbol{\mu}} F(\boldsymbol{\Theta},\boldsymbol{\Psi}) = \frac{1}{G}\sum_g(\mathbf{x}_g+\boldsymbol{\Omega}\mathbf{y}_g)$$
$$\boldsymbol{\Sigma} = \arg\max_{\boldsymbol{\Sigma}} F(\boldsymbol{\Theta},\boldsymbol{\Psi}) = \mathrm{diag}\Big(\mathbf{U}+\mathbf{V}+\mathbf{W}\Big)$$
$$\pi = \arg\max_{\boldsymbol{\pi}} F(\boldsymbol{\Theta},\boldsymbol{\Psi}) = \frac{\sum_{(g,k)|c_{gk}=1}\beta_{gk}}{\sum_{(g,k)|c_{gk}=1}}$$
$$\alpha = \arg\max_{\alpha} F(\boldsymbol{\Theta},\boldsymbol{\Psi}) = \frac{\sum_k \nu_k}{K}$$
(13)

We iterate between the above three steps until we arrive at a maximum of $F(\mathbf{\Theta}, \mathbf{\Psi})$, at which point we examine the β_{gk} targeting probabilities computed from the data and use those to score the candidate interaction between mRNA g and miRNA k.

References

[Attias 1999] Attias, H. Inferring parameters and structure of unobserved variable models by variational Bayes, 21-30. *In* Laskey, K.B., Prade, H., eds., *Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI)*, Morgan Kaufmann Publishers (1999).

[Bishop 2006] Bishop, C. Machine learning and pattern recognition. (New York: Springer-Verlag) (2006).

[Neal and Hinton 1998] Neal, R.M., and Hinton, G.E. A view of the EM algorithm that justifies incremental, sparse, and other variants, 355-368. *In* Jordan, M.I., ed., *Learning in Graphical Models*, MIT Press (1998).