Structured Variational Approximation for Gaussian Assembly Graphs

Christopher Quince

1 The model

We consider an assembly graph comprised of unitig sequennce nodes $v = 1, \dots, V$ and directed edges u, v defining overlaps. We define:

- Counts $x_{v,s}$ for each unitig $v = 1, \dots, V$ in sample $s = 1, \dots, S$
- Paths for strain $g = 1, \dots, G$ defined by $\eta^g_{u,v} \in 0, 1$
- Flow of strain g through v, $\phi^{g+}_v = \sum_{u \in A(v)} \eta^g_{u,v}$ and $\phi^{g-}_v = \sum_{u \in D(v)} \eta^g_{v,u}$ where $A(v)$ is set of ancestors of v and $D(v)$ descendants in the assembly graph
- The following is true $\phi^{g+}_v = \phi^{g-}_v = \phi^g_v$
- Strain coverages $\gamma_{g,s}$
- Unitig lengths L_v
- Unitig bias θ_v
- Source node s and sink node t

Then assume normally distributed counts for each node in each sample giving a joint density for observations and latent variables:

$$
\begin{aligned}
P(\mathbf{X}, \mathbf{\Gamma}, \mathbf{H}, \mathbf{\Theta}) = \prod_{v=1}^{V} \prod_{s=1}^{S} \mathcal{N}(x_{v,s} | L_v \theta_v [\sum_{h=1}^{G} \phi^h_v \gamma_{h,s}], \tau^{-1}) \prod_{h=1}^{G} \prod_{s=1}^{S} P(\gamma_{h,s} | \lambda_h) \\
\cdot \prod_{h=1}^{G} \prod_{v=1}^{V} \left[\phi^{h+}_v = \phi^{h-}_v\right] \left[\phi^{h-}_s = 1\right] \left[\phi^{h+}_t = 1\right] P(\tau) \\
\cdot \prod_{h=1}^{G} P(\lambda_h | \alpha_0, \beta_0) \prod_{v=1}^{V} P(\theta_v | \mu_0, \tau_0) \quad (1)
\end{aligned}
$$

where [] indicates the Iverson bracket evaluating to 1 if the condition is true and zero otherwise. We assume an exponential prior for the $\gamma_{g,s}$ with a rate parameter that is strain dependent, such that:

$$
P(\gamma_{g,s} | \lambda_g) = \lambda_g \exp(-\gamma_{g,s} \lambda_g) \quad (2)
$$

We then place gamma hyper-priors on the λ_g:

$$
P(\lambda_g | \alpha_0, \beta_0) = \frac{\beta_0^{\alpha}}{\Gamma(\alpha_0)} \lambda_g^{\alpha_0 - 1} \exp(-\beta_0 \lambda_g) \quad (3)
$$

and a Gamma prior for the precision:

$$P(\tau|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\tau^{\alpha-1}\exp(-\beta\tau) \tag{4}$$

for the biases θ_v we use a truncated normal prior:

$$\begin{aligned} P(\theta_v|\mu_0,\tau_0) &= \qquad \frac{\sqrt{\frac{\tau_0}{2\pi}}\exp(-\frac{\tau_0}{2}(\theta_v-\mu_0)^2)}{1-\Psi(-\mu_0\sqrt{\tau}_0)} \quad \theta_v >= 0 \\ &= \qquad 0 \quad \theta_v < 0 \end{aligned}$$

where Ψ is the standard normal cumulative distribution.

2 Variational Approximation

We use variational inference to obtain an approximate solution to the posterior distribution of this model (1). In particular because all the distributions are conjugate we can employ CAVI, coordinate ascent variational inference. Our starting point is to assume the following factorisation for the variational approximation:

$$q(\mathbf{X},\mathbf{\Gamma},\mathbf{H}) = \prod_{h=1}^{G} q_h(\{\eta^h_{v,u}\}_{u,v\in A}) \prod_{h=1}^{G}\prod_{s=1}^{S} q_h(\gamma_{h,s}) \prod_{h=1}^{G} q_h(\lambda_h) \prod_{v=1}^{V} q_v(\theta_v) q(\tau) \tag{5}$$

where A are all pairs of nodes in assembly graph. Note that we assumed a fully factorised approximation except for the $\eta^h_{v,u}$, the paths for each strain through the graph. There we assume that the path for each strain forms a separate factor allowing strong correlations between the different elements of the path. Then the mean field update for each set of $\{\eta^h_{v,u}\}_{u,v\in A}$ is derived as:

$$\begin{aligned} \ln q^*_h(\{\eta^h_{v,u}\}_{u,v\in A}) &= \langle \ln P\rangle_{\phi^{h\neq g}_{v\in A},\gamma_{g,s},\theta_v,\lambda_g} \\ &= \ln\left(\prod_{v=1}^{V}\delta_{\phi^{g+}_v,\phi^{g-}_v}\delta_{\phi^{g-}_s,1}\delta_{\phi^{g+}_t,1}\right) \\ &- \left\langle \sum_{v=1}^{V}\sum_{s=1}^{S}\frac{\tau}{2}\left(x_{v,s}-\theta_v L_v[\sum_{h=1}^{G}\phi^h_v\gamma_{h,s}]\right)^2\right\rangle_{\phi^{h\neq g}_{v\in A},\gamma_{g,s}} \end{aligned}$$

Consider the second term only:

$$-\frac{\langle\tau\rangle}{2}\left(-\sum_{v=1}^{V}\sum_{s=1}^{S}2x_{v,s}L_v\langle\theta_v\rangle\langle\gamma_{g,s}\rangle\phi^g_v + L_v^2\langle{\theta_v}^2\rangle\langle(\sum_{h=1}^{G}\phi^h_v\gamma_{h,s})(\sum_{g=1}^{G}\phi^g_v\gamma_{g,s})\rangle)\right)$$

$$-\frac{\langle\tau\rangle}{2}\left(\sum_{v=1}^{V}\sum_{s=1}^{S}\left[-2x_{v,s}\langle\theta_v\rangle L_v\langle\gamma_{g,s}\rangle\phi_v^g+2L_v^2\langle{\theta_v}^2\rangle\sum_{h\neq g}^{G}\langle\phi_v^h\rangle\langle\gamma_{h,s}\rangle\langle\gamma_{g,s}\rangle\phi_v^g+L_v^2\langle{\theta_v}^2\rangle\langle\gamma_{g,s}^2\rangle(\phi_v^g)^2\right]\right)$$

This gives a posterior distribution for the $q_h(\{\eta_{v,u}^h\}$ that is discrete. We solve this together with the constraints
Next we consider the mean field update for the $\gamma_{g,s}$

$$\begin{aligned}\ln q^*(\gamma_{g,s}) &= \langle\ln P\rangle_{\phi_{v\in A}^h,\gamma_{h\neq g,t\neq s}}\\ &= -\left\langle\sum_{v=1}^{V}\frac{\tau}{2}\left(x_{v,s}-\theta_v L_v[\sum_{h=1}^{G}\phi_v^h\gamma_{h,s}]\right)^2\right\rangle_{\phi_{v\in A}^h,\gamma_{h\neq g,t\neq s}}-\frac{\gamma_{g,s}}{\epsilon}\end{aligned}$$

$$\ln q^*(\gamma_{g,s}) = \quad = -\frac{\langle\tau\rangle}{2}\left(\sum_{v=1}^{V}-2x_{v,s}\langle\theta_v\rangle L_v\langle\phi_v^g\rangle\gamma_{g,s}+2\langle\theta_v^2\rangle L_v^2\gamma_{g,s}\langle\phi_v^g\rangle\sum_{h\neq g}\langle\gamma_{h,s}\rangle\langle\phi_v^h\rangle+\langle\theta_v^2\rangle L_v^2\gamma_{g,s}^2\langle[\phi_v^g]^2\rangle\right)-\frac{\gamma_{g,s}}{\epsilon}$$

with the restriction $\gamma_{g,s} > 0$ this gives a truncated Normal distribution for $\gamma_{g,s}$. With mean:

$$\mu_{g,s} = \frac{\sum_v x_{v,s}\theta_v L_v\langle\phi_v^g\rangle-\langle\phi_v^g\rangle\sum_{h\neq g}\langle\gamma_{h,s}\rangle\langle\phi_v^h\rangle\langle\theta_v^2\rangle L_v^2}{\sum_v L_v^2\langle[\phi_v^g]^2\rangle}-\frac{1}{\epsilon\tau_{g,s}} \tag{6}$$

$$\tau_{g,s} = \langle\tau\rangle\sum_v L_v^2\langle[\phi_v^g]^2\rangle \tag{7}$$

$$\tag{8}$$

References

1. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: A review for statisticians. Journal of the American Statistical Association **112**(518), 859–877 (2017). doi:10.1080/01621459.2017.1285773