Appendix to: Inference of population structure from ancient DNA

Tyler A. Joseph[1] and Itsik Pe'er[1,2,3]

[1] Department of Computer Science,
[2] Department of Systems Biology,
[3] Data Science Institute,
Columbia University, New York, NY 10027, USA
{tjoseph, itsik}@cs.columbia.edu

A Stochastic Variational Inference

In this appendix we derive the inference algorithm for stochastic variational inference under the Dystruct model. We first derive the traditional coordinate ascent updates, then show how we can modify these updates for stochastic optimization. Finally, we extend the algorithm for missing data.

A.1 Computing The ELBO

The ELBO is given by

$$\begin{aligned}
L &= \mathbb{E}_q[\log p(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D}, \boldsymbol{x}_{1:D,1:L})] - \mathbb{E}_q[\log q(\boldsymbol{\beta}_{1:K,1:L}, \boldsymbol{\theta}_{1:D})] && (1)\\
&= \sum_{t=1}^{T} \sum_{k=1}^{K} \sum_{l=1}^{L} \mathbb{E}_q[\log p(\beta_{kl}[t] \big| \beta_{kl}[t-1])] && (2)\\
&\quad + \sum_{d=1}^{D} \mathbb{E}_q[\log p(\boldsymbol{\theta}_d)] && (3)\\
&\quad + \sum_{d=1}^{D} \sum_{l=1}^{L} \mathbb{E}_q[\log p(x_{dl} | t_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K,l})] && (4)\\
&\quad - \sum_{k=1}^{K} \sum_{l=1}^{L} \mathbb{E}_q[\log q(\boldsymbol{\beta}_{kl} \big| \hat{\boldsymbol{\beta}}_{kl})] && (5)\\
&\quad - \sum_{d=1}^{D} \mathbb{E}_q[\log q(\boldsymbol{\theta}_d \big| \hat{\boldsymbol{\theta}}_d)] && (6)
\end{aligned}$$

The ELBO as written does not have a closed form due to the log sum terms that appear in (4): $\mathbb{E}_q[\log p(x_{dl} \big| t_d, \boldsymbol{\theta}_d, \boldsymbol{\beta}_{1:K,l})] = \mathbb{E}_q[\log \text{Binomial}(2, \sum_k \beta_{kl}[t_d]\theta_{dk})]$:

$$x_{dl}\, \mathbb{E}_q\left[\log\left(\sum_k \theta_{dk}\beta_{kl}[t_d]\right)\right] + (2 - x_{dl})\, \mathbb{E}_q\left[\log\left(1 - \sum_k \theta_{dk}\beta_{kl}[t_d]\right)\right] \tag{7}$$

Following [4], we optimize a surrogate lower bound by introducing auxiliary variational parameters $\boldsymbol{\phi}_{dl} = (\phi_{dl}[1], ..., \phi_{dl}[K])$ and $\boldsymbol{\zeta}_{dl} = (\zeta_{dl}[1], ..., \zeta_{dl}[K])$ whose vector components sums to 1. An application of Jensen's inequality shows

$$\log\left(\sum_k \theta_{dk}\beta_{kl}[t_d]\right) \geq \sum_k \phi_{dl}[k] \log\left(\frac{\theta_{dk}\beta_{kl}[t_d]}{\phi_{dl}[k]}\right) \tag{8}$$

$$\log\left(1 - \sum_k \theta_{dk}\beta_{kl}[t_d]\right) \geq \sum_k \zeta_{dl}[t_d] \log\left(\frac{\theta_{dk}(1 - \beta_{kl}[t_d])}{\zeta_{dl}[k]}\right) \tag{9}$$

so we still maintain a lower bound on the log likelihood. The auxiliary parameters are optimized to provide a tight lower bound. Fixing all other parameters, the constrained optimization problem can be solved using an application of Lagrange multipliers

$$\phi_{dl}[k] \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log \beta_{kl}[t_d]]\} \tag{10}$$

$$\zeta_{dl}[k] \propto \exp\{\mathbb{E}_q[\log \theta_{dk}] + \mathbb{E}_q[\log(1 - \beta_{kl}[t_d])]\} \tag{11}$$

The first term in both equations is an expectation of a sufficient statistic, and therefore has a closed form: $\mathbb{E}_q[\log \theta_{dk}] = \Psi(\hat{\theta}_{dk}) - \Psi(\sum_k \hat{\theta}_{dk})$; Ψ is the Digamma function. The two expectations in the second terms can be approximated by taking second order Taylor expansions around the marginal means of $\beta_{kl}[t_d]$ and $\widetilde{m}_{kl}[t]$:

$$\mathbb{E}_q[\log \beta_{kl}[t_d]] \approx \log \widetilde{m}_{kl}[t_d] - \frac{\widetilde{v}_{kl}[t_d]}{2\widetilde{m}_{kl}[t_d]^2} \tag{12}$$

$$\mathbb{E}_q[\log(1 - \beta_{kl}[t_d])] \approx \log(1 - \widetilde{m}_{kl}[t_d]) \tag{13}$$

A.2 Optimizing Ancestry Proportions

Note that the $q(\boldsymbol{\theta}_d | \hat{\boldsymbol{\theta}}_d)$ satisfy the mean field assumption - the $\boldsymbol{\theta}_d$ in the variational posterior are independent. Therefore they have optimal coordinate ascent updates of the form

$$q^*(\boldsymbol{\theta}_d) \propto \exp\{\mathbb{E}_q[\log p(\boldsymbol{\theta}_d | \boldsymbol{\beta}_{kl}[t_d], \boldsymbol{x}_{d,1:L})]\} \tag{14}$$

where we have used several conditional independencies to simplify the complete conditional of $\boldsymbol{\theta}_d$. Using the surrogate lower bound in the ELBO gives the optimal update

$$\hat{\theta}_{dk} = \alpha_k + \sum_{l=1}^{L} x_{dl}\phi_{dl}[k] + (2 - x_{dl})\zeta_{dl}[k] \tag{15}$$

matching the expression in [4].

A.3 Optimizing allele frequencies

In variational Kalman filtering, the variational distribution for each $\beta_{kl}[t]$ is given by

$$\beta_{kl}[t] \sim \text{Normal}(\widetilde{m}_{kl}[t], \widetilde{v}_{kl}[t]) \tag{16}$$

where the mean and variance are the marginal means and posteriors given by the Kalman filtering and smoothing equations. Following the notation in [3], the forward (filtered) means and variances are given by

$$m_{kl}[t] = \frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2} m_{kl}[t-1] + \left(1 - \frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) \hat{\beta}_{kl}[t] \tag{17}$$

$$v_{kl}[t] = \left(\frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) (v_{kl}[t-1] + \sigma_k^2[t]) \tag{18}$$

where $\sigma_k^2[t] := \frac{\Delta g[t]}{12 N_k}$. The initial conditions are $m_{kl}[0] = \beta_{kl}[0]$. The marginal (smoothed) means and variances are

$$\widetilde{m}_{kl}[t] = \left(\frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) m_{kl}[t] + \left(1 - \frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \widetilde{m}_{kl}[t+1] \tag{19}$$

$$\widetilde{v}_{kl}[t] = v_{kl}[t] + \left(\frac{v_{kl}[t]}{v_{kl}[t] + \sigma_k^2[t]}\right)^2 (\widetilde{v}_{kl}[t+1] - v_{kl}[t] - \sigma_k^2[t+1]) \tag{20}$$

with initial conditions $\widetilde{m}_{kl}[T] = m_{kl}[T]$ and $\widetilde{v}_{kl}[T] = v_{kl}[T]$. The variational parameters $\hat{\beta}_{kl}[t]$ are optimized with respect to the ELBO, hence we need the partial derivatives of the marginal means

$\widetilde{m}_{kl}[t]$ with respect to $\hat{\beta}_{kl}[t]$. These can be obtained using the forward-backward recurrence as in [3]. We will show the recurrence for initial frequencies $\beta_{kl}[0]$, which are not maximized in [3], and note that the other partial derivations can be obtained similarly. The recurrence is

$$\frac{\partial m_{kl}[t]}{\partial \beta_{kl}[0]} = \left(\frac{\nu^2}{v_{kl}[t-1] + \sigma_k^2[t] + \nu^2}\right) \frac{\partial m_{kl}[t-1]}{\beta_{kl}[0]} \tag{21}$$

$$\frac{\partial \widetilde{m}_{kl}[t]}{\beta_{kl}[0]} = \left(\frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \frac{\partial m_{kl}[t]}{\partial \beta_{kl}[0]} + \left(1 - \frac{\sigma_k^2[t]}{v_{kl}[t] + \sigma_k^2[t]}\right) \frac{\partial \widetilde{m}_{kl}[t+1]}{\beta_{kl}[0]} \tag{22}$$

We optimize the $\hat{\boldsymbol{\beta}}_{kl}$ with respect to a single locus in a single population at time using a conjugate gradient algorithm, constraining the parameters to lie in the interval $(0, 1)$. The terms in the ELBO with respect to locus l in population k are

$$\begin{aligned}
L_* = &\sum_{t=1}^{T} \mathbb{E}_q[\log p(\beta_{kl}[t] \big| \beta_{kl}[t-1])] - \mathbb{E}_q[\log q(\beta_{kl}[t] \big| \widetilde{m}_{kl}[t], \widetilde{v}_{kl}[t])] \qquad (23)\\
&+ \sum_{t=1}^{T} \sum_{d:t_d=t} \mathbb{E}_q[\log p(x_{dl} \big| t_d, \beta_{kl}[t_d], \boldsymbol{\theta}_d)]\\
\geq &-\frac{T}{2}\log 2\pi - \frac{1}{2}\sum_{t=1}^{T} \log \sigma_k^2[t] - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]} \mathbb{E}_q[(\beta_{kl}[t] - \beta_{kl}[t-1])^2] \qquad (24)\\
&+ \frac{T}{2}\log 2\pi + \frac{T}{2} + \frac{1}{2}\sum_{t=1}^{T} \log \widetilde{v}_{kl}[t]\\
&+ \sum_{t=1}^{T} \sum_{d:t_d=t} x_{dl}\phi_{dl}[k] \left(\log \widetilde{m}_{kl}[t] - \frac{\widetilde{v}_{kl}[t]}{2\widetilde{m}_{kl}[t]^2}\right) + (2 - x_{dl})\zeta_{dl}[k] \log(1 - \widetilde{m}_{kl}[t])\\
= &\frac{T}{2} - \frac{1}{2}\sum_{t=1}^{T}(\log \sigma_k^2[t] - \log \widetilde{v}_{kl}[t]) - \frac{1}{2}\sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]}(\widetilde{m}_{kl}[t] - \widetilde{m}_{kl}[t-1])^2 - \frac{\widetilde{v}_{kl}[t]}{\sigma_k^2[t]} - \frac{\widetilde{v}_{kl}[t-1]}{\sigma_k^2[t-1]} \qquad (25)\\
&+ \sum_{t=1}^{T} \sum_{d:t_d=t} x_{dl}\phi_{dl}[k] \left(\log \widetilde{m}_{kl}[t] - \frac{\widetilde{v}_{[}kl][t]}{2\widetilde{m}_{kl}[t]^2}\right) + (2 - x_{dl})\zeta_{dl}[k] \log(1 - \widetilde{m}_{kl}[t])
\end{aligned}$$

where we define $\widetilde{v}_{kl}[0] = 0$, $\sigma_k^2[0] = 1$, and $m_{kl}[0] = \widetilde{m}_{kl}[0] = \beta_{kl}[0]$ for notational convenience. Taking partial derivatives with respect to the pseudo-outputs gives us

$$\begin{aligned}
\frac{\partial L_*}{\partial \hat{\beta}_{kl}[s]} = &-\sum_{t=1}^{T} \frac{1}{\sigma_k^2[t]}(\widetilde{m}_{kl}[t] - \widetilde{m}_{kl}[t-1]) \left(\frac{\partial \widetilde{m}_{kl}[t]}{\partial \hat{\beta}_{kl}[s]} - \frac{\partial \widetilde{m}_{kl}[t-1]}{\partial \hat{\beta}_{kl}[s]}\right) \qquad (26)\\
&+ \frac{\partial \widetilde{m}_{kl}[t]}{\partial \hat{\beta}_{kl}[s]} \sum_{d:t_d=t} x_{dl}\phi_{dl}[k] \left(\frac{1}{\widetilde{m}_{kl}[t]} + \frac{\widetilde{v}_k[t]}{\widetilde{m}_{kl}[t]^3}\right) + (2 - x_{dl})\zeta_{dl}[k] \frac{1}{(\widetilde{m}_{kl}[t] - 1)}
\end{aligned}$$

The full algorithm iterates between optimizing the local parameters $\hat{\boldsymbol{\beta}}_{kl}$, $\phi_{dl}[k]$, and $\zeta_{dl}[k]$ for each locus in each individual in each population using (10), (11), and (26), then updating global parameters $\hat{\boldsymbol{\theta}}_d$ according to (15) until convergence.

A.4 Inference Algorithm

We can perform stochastic variational inference through a slight modification to the coordinate ascent algorithm presented above [2, 5]. Stochastic variational inference computes noisy estimates of the optimal global parameters by stochastically subsampling data points, and using the optimal local parameters to update the global parameters. The optimal global parameters are a weighted average of the previous global parameters, with the newly computed global parameters. Following [4], the $n+1$ stochastic variational inference update for the global parameters $\hat{\boldsymbol{\theta}}_d$ is

$$\hat{\theta}_{dk}^{n+1} = (1-\epsilon_n)\hat{\theta}_{dk}^{n} + \alpha_k + \epsilon_n L \left(x_{dl}\phi_{dl}[k] + (2-x_{dl})\zeta_{dl}[k] \right) \tag{27}$$

where ϵ_n is the step size for iteration n and L is the number of loci. Provided the step size meets certain criteria the algorithm is guaranteed to converge. See [5] or [2] for more details. We picked a step size of $\epsilon_n = (1+n)^{-0.5}$ for the first 10000 iterations, and $\epsilon_n = (n-7825)^{-0.6}$ for the remaining iterations.

Algorithm 1 Dystruct inference algorithm

```
Input: Genotypes $\boldsymbol{x}_{1:D,1:L}$; Sample Times $t_d$; Population Size $N_k = N$ for all populations.
while $\hat{\boldsymbol{\theta}}_d$ have not converged do
    Pick $l \sim \text{Uniform}(1, L)$
    while $\boldsymbol{\phi}_{dl}$ and $\boldsymbol{\zeta}_{dl}$ have not converged do
        Update auxiliary parameters $\boldsymbol{\phi}_{dl}$ and $\boldsymbol{\zeta}_{dl}$ for $d = 1, 2, ..., D$ according to (10) and (11).
        Update allele frequency parameters $\hat{\boldsymbol{\beta}}_{kl}$ for $k = 1, 2, ..., K$ using the numerical optimization routine described in section A.3.
    end while
    Update global parameters $\hat{\boldsymbol{\theta}}_d$ for $d = 1, 2, ..., D$ according to (27)
end while
```

A.5 Extensions to missing data

The above algorithm only holds for complete data. A small modification is required for missing data, where not every sample has an observed genotype at every locus. Rather than a single global step size ϵ_t, we maintain a step size for every individual ϵ_{n_d} where n_d is the number of iterations for individual d. When a locus is subsampled, we only update global ancestry estimates for individuals with observed genotypes at that locus, and the step size for those individuals. We further replace the parameter L with L_d, the number of loci for each observed in each individual.

References

1. Alexander, D.H., Novembre, J., Lange, K.: Fast model-based estimation of ancestry in unrelated individuals. Genome Res 19(9), 1655–1664 (2009)
2. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J Am Stat Assoc 112(518), 859–877 (2017)
3. Blei, D.M., Lafferty, J.D.: Dynamic topic models. In: Proc Int Conf Mach Learn. pp. 113–120. ACM (2006)
4. Gopalan, P., Hao, W., Blei, D.M., Storey, J.D.: Scaling probabilistic models of genetic variation to millions of humans. Nat Genet 48(12) (2016)
5. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.W.: Stochastic variational inference. J Mach Learn Res 14(1), 1303–1347 (2013)
6. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics 155(2), 945–959 (2000)