Homework: Variational inference for population structure mixture models

1 Background

In a large number of biology applications, the data consist of measurements from a number of organisms, and it is important to understand the "population structure" exhibited by these organisms. Organisms tend to segregate into populations, such that individuals within a given population interbreed commonly, while breeding between populations is much less common. This causes the data to tend to fall into clusters corresponding to these populations. It is important to infer these clusters and take them into account for many types of analysis. Failure to account for population structure can lead to misleading results, due to Simpson's paradox. The standard way to infer population structure from genotype data is to use what is called an "admixture model". We will consider a simplification based on an ordinary mixture model. The article introducing these models (Pritchard et al., 2000) is one of the most highly cited statistics papers of all time, with 16,498 citations as of this writing, according to Google Scholar.

2 Data

We will consider data from Lorenzen and Siegismund (2004) and Lorenzen et al. (2006), consisting of genotypes from $n = 216$ common impala and black-faced impala from Southern Africa (see Figure 1). The common impala is widespread throughout the eastern part of Southern Africa, whereas the black-faced impala is an endangered subspecies that is localized to a small region in the western part. These researchers were interested in understanding the genetic diversity of these animals, in order to help save the black-faced impala from extinction.

For each animal $j = 1, \ldots, n$, its genotype was determined at $L = 8$ loci (i.e., 8 locations on the genome). At each locus $\ell = 1, \ldots, 8$, the genotype of animal j consists of two allele copies, $x_{j\ell 1}, x_{j\ell 2} \in \{1, \ldots, V_\ell\}$ (since each animal has two copies of each chromosome, one from each parent). Here, V_ℓ is the number of different variants of allele copy that can occur at locus ℓ, and for this data set, $V = (15, 13, 6, 6, 9, 14, 16, 9)$. For example, Table 1 shows the data for animal $j = 86$. Missing entries are indicated by -1. The file "x.txt" contains the genotype data for all $n = 216$ animals.

Jeffrey W. Miller (2016). *Lecture Notes on Advanced Stochastic Modeling.* Duke University, Durham, NC.

Figure 1: Left: Male common impala. Middle: Male black-faced impala. Right: Female and offspring black-faced impala.[1]

Table 1: Genotype of animal $j = 86$

Locus ℓ	1		2		3		4		5		6		7		8	
Allele copy c	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Variant observed, $x_{j\ell c}$	5	3	2	2	-1	-1	2	2	1	2	3	6	3	2	1	1

3 Model

Consider the following model. Suppose there are k mixture components, representing k different populations. The component weights are $w = (w_1, \ldots, w_k)$, where $w_i \geq 0$ and $\sum_{i=1}^{k} w_i = 1$. The component parameters are $\alpha = (\alpha_1, \ldots, \alpha_k)$, where for each component $i = 1, \ldots, k$,

$$\alpha_i = \begin{bmatrix} \alpha_{i1}(1), \ldots, \alpha_{i1}(V_1) \\ \vdots \\ \alpha_{i\ell}(1), \ldots, \alpha_{i\ell}(V_\ell) \\ \vdots \\ \alpha_{iL}(1), \ldots, \alpha_{iL}(V_L) \end{bmatrix},$$

with $\alpha_{i\ell}(v)$ being the probability of observing variant v at locus ℓ for an animal in population i. Note that this is a "ragged matrix", i.e., the rows have different lengths. Each row $\alpha_{i\ell}$ is a probability vector, in other words, $\alpha_{i\ell}(v) \geq 0$ and $\sum_{v=1}^{V_\ell} \alpha_{i\ell}(v) = 1$.

The data is modeled as

$$\begin{aligned} &Z_j | w \sim \text{Categorical}(w) \\ &X_{j\ell 1}, X_{j\ell 2} \mid \alpha,\, Z_j = i \sim \text{Categorical}(\alpha_{i\ell}) \text{ independently for } \ell = 1, \ldots, L, \end{aligned}$$

independently for each $j = 1, \ldots, n$.[2] See Figure 2 for the graphical model. We will use uniform priors on w and on $\alpha_{i\ell}$ for all i, ℓ.

[1] Image credits: Common impala, Filip Lachowski, CC BY-SA 2.0. Black-faced impala, Yathin S Krishnappa, CC BY-SA 3.0. Female and offspring, Yathin S Krishnappa, CC BY-SA 3.0.

[2] Technically, the allele counts at each locus should be Multinomial$(2, \alpha_{i\ell})$, since the order of the two allele copies is undetermined. However, to keep the assignment from getting too complicated, we will use the simpler Categorical model.

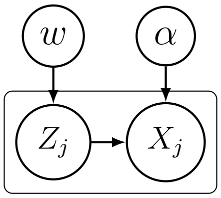

Figure 2: Directed graphical model.

In this homework, you will construct a variational approximation to the posterior on z, w, α for this model. The target distribution is

$$\begin{aligned}\pi(z,w,\alpha) = p(z,w,\alpha|x) &\propto p(x,z,w,\alpha) \propto p(x|z,\alpha)p(z|w)\\ &= \prod_{j=1}^{n}\Big(\prod_{\ell=1}^{L}\prod_{c=1}^{2}\alpha_{z_j\ell}(x_{j\ell c})\Big)w_{z_j}.\end{aligned}$$

Taking the log and using indicator functions to simplify the dependence on the parameters, we get

$$\log\pi(z,w,\alpha) = \sum_{j=1}^{n}\sum_{i=1}^{k}\mathbb{1}(z_j = i)\Big(\log w_i + \sum_{\ell=1}^{L}\sum_{c=1}^{2}\sum_{v=1}^{V_\ell}\mathbb{1}(x_{j\ell c} = v)\log\alpha_{i\ell}(v)\Big) + \text{const.}$$

To handle the missing data, let's assume the data collection process is ignorable (since the model is already complicated enough for this homework exercise). With this assumption, the same exact expression as above can be used for $\log\pi(z,w,\alpha)$, for roughly the following reason: in the sum over $v = 1,\ldots,V_\ell$, the indicator $\mathbb{1}(x_{j\ell c} = v)$ will always be zero if $x_{j\ell c}$ is missing, because -1 is never equal to v; therefore any missing data will not factor into the likelihood. Note, however, that you will need to be careful in manipulating this expression, since this means that $\sum_{v=1}^{V_\ell}\mathbb{1}(x_{j\ell c} = v)\log\alpha_{i\ell}(v) \neq \log\alpha_{i\ell}(x_{j\ell c})$ if $x_{j\ell c}$ is missing, because the left-hand side is zero and $\alpha_{i\ell}(-1)$ is undefined.

4 Exercises

1. Derive the variational inference algorithm for this model based on an approximating distribution of the form $q(z,w,\alpha) = q(z)q(w,\alpha)$. Hint: Similarly to the mixture model we considered in class, for the $q(z)$ update, the algorithm will involve computing $r_j(i) = \mathbb{P}_q(Z_j = i)$ (but with a different formula than before), and for the $q(w,\alpha)$ update, it will involve computing $R_i = \sum_{j=1}^{n} r_j(i)$ as well as some other quantities $S_{i\ell}(v)$ that you will need to determine.

2. Implement the algorithm using a random initialization. For the convergence criterion, stop when the root-mean-square difference between r^{new} and r^{old} is less than 10^{-10}, that is, when

$$\Big(\frac{1}{nk}\sum_{j=1}^{n}\sum_{i=1}^{k}\big(r_j^{\text{new}}(i) - r_j^{\text{old}}(i)\big)^2\Big)^{1/2} < 10^{-10}.$$

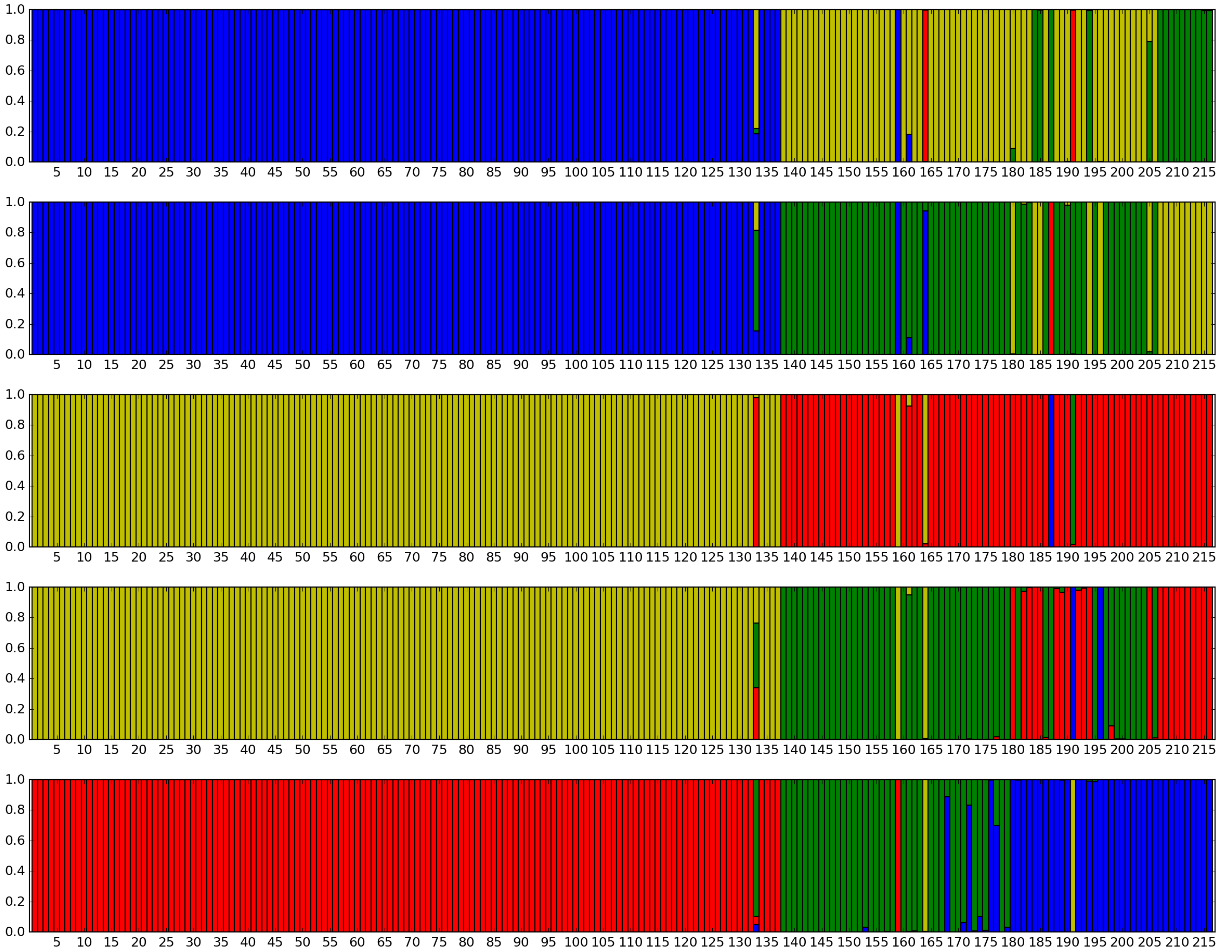

Figure 3: Example results from five runs with different random initializations.

3. Now, with $k = 4$ components, run the algorithm five times on the data in "x.txt", using a different random initialization each time. For each of these five runs, report the number of iterations until convergence, and plot the final values of $r_j(i)$ in a stacked bar plot as shown in Figure 3. Discuss the interpretation of these plots.

4. Table 2 indicates the subspecies ("C" for common impala or "B" for black-faced impala) and the region of origin for each animal. Do your results appear to make sense, at least roughly, in light of this additional information? Discuss. Are there any individual animals who are consistently clustered differently than other members of their subspecies and region? Give the indices j of two or three animals like this, and discuss possible explanations.

Note that the algorithm does not converge to the same thing every time. For example, sometimes, two given animals will be assigned to the same component with very high probability in one run, but assigned to two different components with very high probability in another run. This may be because these kinds of variational approximations tend to underestimate uncertainty (in other words, they tend to be "overconfident").

Table 2: Subspecies and region for each animal.

Animal j	Subspecies	Region code
1-15	B	KA
16-31	B	OM
32-64	B	OL
65-98	B	HA
99-127	B	NA
128-137	B	ON
138-147	C	CH
148-158	C	SH
159-163	C	KAF
164-169	C	LU
170-181	C	SE
182-194	C	BU
195-206	C	IM
207-216	C	SA

References

Eline D Lorenzen and Hans R Siegismund. No suggestion of hybridization between the vulnerable black-faced impala (Aepyceros melampus petersi) and the common impala (A. m. melampus) in Etosha National Park, Namibia. *Molecular Ecology*, 13(10):3007–3019, 2004.

Eline D Lorenzen, Peter Arctander, and Hans R Siegismund. Regional genetic structuring and evolutionary history of the impala Aepyceros melampus. *Journal of Heredity*, 97(2): 119–132, 2006.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.