My Hubris on Display

Nitpicking the pedagogical case study of two Stanford statisticians

Jared Cummings

Fall 2021

Data

- Genotype data of 197 individuals at $m = 100$ *well-space* single-nucleotide polymorphism (SNP)
- 29 individuals with missing values *implies* $n = 168$
- Genotype of individual i at locus m, G_{im}, is three level factor
  - two wild type genes: `AA` → 0
  - one wild type, one mutation: `Aa` → 1
  - two mutations: `aa` → 2
- self-reported ethnicity reported

Subject	SNP_1	SNP_2	SNP_3	...	SNP_{97}	SNP_{98}	SNP_{99}	SNP_{100}
NA10852	1	1	0	...	1	1	0	0
NA12239	1	1	0	...	1	1	0	0
NA19072	0	0	0	...	0	0	0	0
NA19247	0	0	2	...	0	0	0	2
NA20126	2	0	0	...	2	0	0	0
NA18868	0	0	1	...	0	0	0	1

Table: A subset of genotype data on 197 individuals, each with measurements at 100 SNPs. Each individuals `ethnicity` is known to be one of `Japanese`, `African`, `African American`, and `European`.

Motivating model

- Efron and Hastie (2016) propose the model

$$\begin{aligned} Z_{im}^{(c)} &\sim \text{Multi}(1, Q_i) \\ X_{im}^{(c)} | Z_{im}^{(c)} = j &\sim \text{Bi}(1, P_{jm}) \\ Q_i &\sim \text{Dir}(1,1,1) \\ P_{jm} &\sim \text{Dir}(1,1). \end{aligned} \tag{1}$$

  where i indexes individuals, m indexes loci, c indexes gene copies
- X_{im} is derived from G_{im} by
  - $G_{im} = 0 \implies X_{im} = (0,0)$
  - $G_{im} = 2 \implies X_{im} = (1,1)$
  - $G_{im} = 1 \implies X_{im} = (0,1) \vee X_{im} = (1,0)$
- Performs soft clustering of individuals on $\mathcal{S}_3$ with Q_i as coordinates

Criticism

Efron and Hastie's model assumes more information than the data G_{im} provide. Three cases:

1. $\Pr(G_{im} = 0) = \Pr(X_{im} = (0,0))$
2. $\Pr(G_{im} = 2) = \Pr(X_{im} = (1,1))$
3. $\Pr(G_{im} = 1) = \Pr(X_{im} = (0,1)) + \Pr(X_{im} = (1,0))$

Questions

What are the consequences of the model proposed by Efron and Hastie on estimating population admixture of `African Americans`...

- ...as it relates to data representation?
- ...as it relates to the concentration parameters on Q_i?

Proposed models

For effect of concentration parameter

$$\begin{aligned} Y_{im}|Q_i &\sim \text{Multi}(1, Q_i) \\ G_{im}|Y_{im}=j, R_{jm} &\sim \text{Multi}(1, R_{jm}) \\ Q_i|\lambda_1,\lambda_2,\lambda_3 &\sim \text{Dir}(\lambda_1,\lambda_2,\lambda_3) \\ \lambda_k &\sim \text{Gam}(2,2) \\ R_{jm} &\sim \text{Dir}(1,1,1). \end{aligned} \tag{2}$$

For effect of data representation, same as above, but

$$\begin{aligned} Q_i|\lambda_1,\lambda_2,\lambda_3 &\sim \text{Dir}(\lambda_1,\lambda_2,\lambda_3) \\ \lambda_k &= 1 \end{aligned} \tag{3}$$

Model fitting

- All three models benefit from conjugacy – Gibbs sampling used
- Model (2) uses Metropolis algorithm (3 univariate Gaussian random walks) for the posterior of hyperparameters λ_k
- More than 10 parameters are being estimated

Table: Number of estimated parameters and effective number of parameters for the three models. The majority are the latent $Z_{im}^{(c)}$ and Y_{im}. Interest lies in the 168 Q_i (504 values total described fully by 336).

	p_{actual}	p_{WAIC}
Model (1)	34,704	2513.1
Model (2)	18,207	871.9
Model (3)	18,204	2058.3

- Ordering of variables everywhere j indexes can be permuted between chains and model fits

Convergence

1. Run models for 10k iterations
2. compute Raftery-Lewis diagnostic
3. rerun at recommended settings
4. Assess convergence with Geweke diagnostic

Table: Settings for single chain convergence recommended by the Raftery-Lewis diagnostic. All models were refit using the recommended settings. The discrepancy between ”by hand” implementations and those from NIMBLE, which automatically selects an algorithm or set of algorithms to use in estimation beraha2021. Though it is likely that NIMBLE is preferring Gibbs sampling here, it is difficult to confirm that to be true. More advanced applications of NIMBLE overcome this uncertainty by allowing the user to directly specify the algorithm(s) used.

Implementation	Model	Thinning	Warm-up samples	Post-warm-up draws
”By hand”	(1)	92	490	344632
	(2)	18	40	67428
	(3)	27	84	101142
NIMBLE	(1)	33	126	123618
	(2)	16	66	59936
	(3)	39	154	146094

Convergence (cont.)

- Geweke diagnostics appear to conform well with standard normal asymptotic behavior at convergence

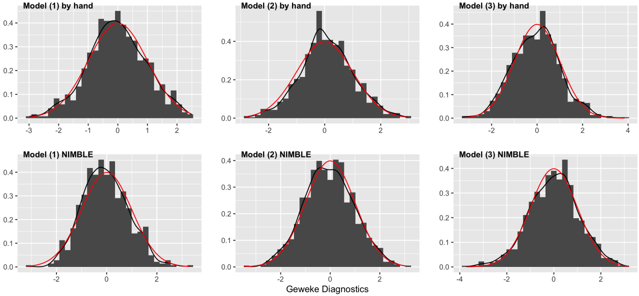


Figure: Geweke diagnostics for the six model fits given in histogram and KDE (black line) against a standard normal curve (red line). At convergence, Geweke diagnostics are asymptotically N(0,1). Slight bias exists in several of the displayed plots, but none of the plots appear heavy tailed.

Comparison of implementations

- Problem: Parameter ordering and the number of parameters makes full comparison impractical
- Solution: Let T_i represent the ethnicity of individual i as recorded in the data

$$\phi_t = \frac{1}{n_t} \sum_{i:T_i=t} \| T_i - \mu_t \|$$

investigate conformity of ϕ_t

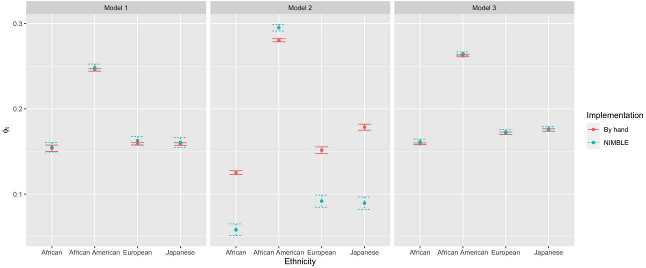


Figure: Comparison of implementations of the three models through the summary value ϕ_t with Monte Carlo error assessed. Both implementations fit similarly when estimating models 1 and 3. Model 2 displays non-trivial differences between implementations. It is interesting to note that between models 1 and 3, in all cases the point estimates of ϕ_t are lower for the model suggested by Efron and Hastie than for a model that does not reduce uncertainty by arbitrary data preprocessing. These estimates can be interpreted as strength of clustering, with lower values being associated with tighter clustering.

Conformity of implementations (cont.)

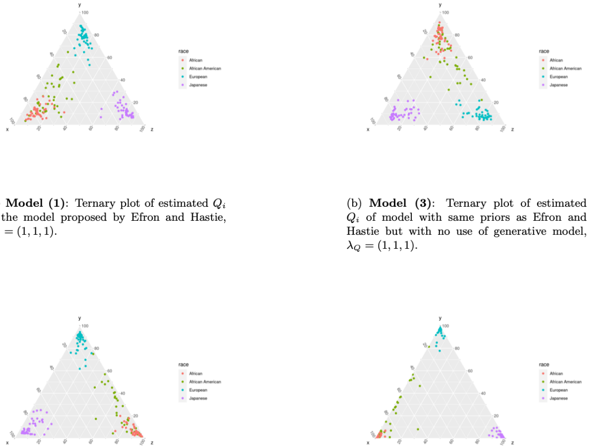
Table: Estimates of concentration hyperparameters. Though inconclusive, the differences in estimates are likely due to differences in estimation procedures as they lead to similar expectations of the elements of Q_i as shown in Table 5.

Implementation	Quantity	Estimate	95% CrI LB	95% CrI UB	95% MC LB	95% MC UB
By hand	λ_1	0.261	0.190	0.346	0.260	0.261
	λ_2	0.374	0.284	0.471	0.373	0.374
	λ_3	0.508	0.418	0.611	0.507	0.508
NIMBLE	λ_1	0.102	0.057	0.227	0.099	0.105
	λ_2	0.072	0.037	0.161	0.068	0.077
	λ_3	0.059	0.029	0.150	0.057	0.062

Table: Estimates of the expected value of the elements of Q_i a priori. Though the estimates of the two implementations are not equivalent up to Monte Carlo error, there are certainly strong similarities in reported values. Apparent differences have a likely source in the different estimated values of the concentration parameters reported in Table 4. Note that the estimates in this table provide a good example of the ordering issue discussed previously.

Implementation	Quantity	Estimate	95% MC LB	95% MC UB	95% CrI LB	95% CrI UB
By hand	$Q_i[1]$	0.228	0.227	0.229	0.171	0.299
	$Q_i[2]$	0.327	0.327	0.327	0.258	0.394
	$Q_i[3]$	0.445	0.445	0.445	0.375	0.513
NIMBLE	$Q_i[1]$	0.450	0.450	0.451	0.370	0.524
	$Q_i[2]$	0.306	0.306	0.306	0.242	0.374
	$Q_i[3]$	0.243	0.243	0.244	0.187	0.316

Results and Sensitivity Analysis



(a) **Model (1)**: Ternary plot of estimated Q_i of the model proposed by Efron and Hastie, $\lambda_Q = (1, 1, 1)$.

(b) **Model (3)**: Ternary plot of estimated Q_i of model with same priors as Efron and Hastie but with no use of generative model, $\lambda_Q = (1, 1, 1)$.

(c) **Model (3) fit by hand**: Ternary plot of estimated Q_i of model with independent gamma priors on λ_Q, $\hat{\lambda}_Q^{\text{MSE}} = (0.261, 0.374, 0.508)$.

(d) **Model (3) fit by NIMBLE**: Ternary plot of estimated Q_i of model with independent gamma priors on λ_Q, $\hat{\lambda}_Q^{\text{MSE}} = (0.102, 0.072, 0.059)$.

Non-Bayesian Method

- Latent class models allow estimation of an analog to Q_i; EM algorithm + log-linear model
- Estimation issues when contingency table is sparse/cell counts are low
- (Highly subjective) variable selection can improve estimation and find meaningful clusters

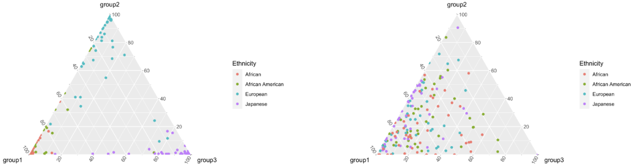


Figure: Estimation of admixture using a latent class model fit with 9 carefully selected SNP loci (left) and all SNP loci with empirical variance greater than 0 (right). Poor fit is to be expected with a sparse contingency table like the kind this data creates. Better clustering than this can be achieved with thoughtful variable selection, though this is a highly subjective process and the Bayesian method requires less input from the modeler.