Genetic Association Studies in the Presence of Population Structure and Admixture

Purushottam W. Laud and Nicholas M. Pajewski

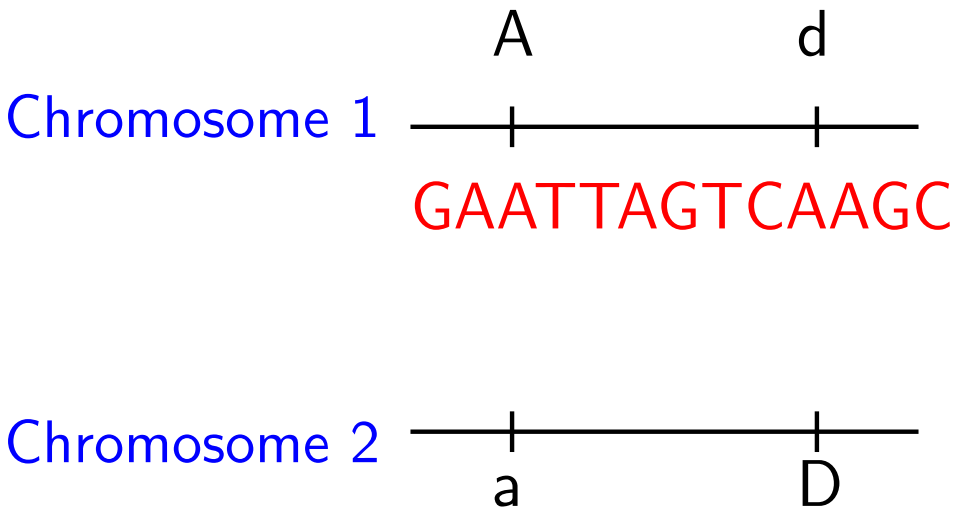
Division of Biostatistics
Department of Population Health
Medical College of Wisconsin

A presentation in the BNP Workshop, INI, Cambridge, August 7, 2007

Outline

- Some genetics background
- Impact of structure on association studies
- Remedies
- A DPM model
- Clustering success
- Estimation success
- A model for admixture
- Detection of admixed individuals

Some Genetics Terminology



- Chromosome
- Locus / Polymorphism
- Allele
- Genotype
- Two alleles **A** and **a** yield three potential genotypes at each locus: **AA**, **aA**/**Aa**, or **aa**
- Phenotype

Propagation, Inheritance and Diversity

- In a pair of homologous chromosomes, one is inherited from the mother and the other from the father
- Recombination can result in alleles that differ from those of parents
- Mechanisms leading to variability in the frequency of genetic variants in human populations
  - Positive Selection
  - Ancestral Admixture
  - Genetic Drift
- Allele frequency assumptions
  - Hardy-Weinberg Equilibrium - Chromosome independence
  - Linkage Disequilibrium - Locus to locus independence

Wright's Fixation Index

- Wright's F_{ST} - Measures the divergence between populations relative to the overall level of genetic diversity within a species (Wright 1951)
- Simulation mechanism {Balding & Nichols (1995)} :
  - Let p denote "ancestral" allele frequency at a locus. This could be drawn from, say, Uniform(0.10,0.90) or fixed
  - Then, π - the allele frequency at this locus is simulated as

$$\pi \sim \text{Beta } \left(rp, r(1-p)\right), \quad r = \frac{(1-F_{ST})}{F_{ST}}$$

  -

$$E(\pi) = p, \; Var(\pi) = p(1-p)F_{ST}$$

Common Study Designs

Main goal is to determine if a phenotype is associated with a genotype at some locus or a combination of loci

- Phenotype of interest can be quantitative or binary
- Sampling design can be cross-sectional or case-control
- Disease presence/absence, sample of cases and sample of non-cases leads to 2×2 table with column totals fixed
- Generally, traditional analyses use linear models and contingency tables

Population Structure and Impact on Association

- If sampling from a homogeneous population with stable allele frequencies at loci of interest, traditional methods can appropriately determine association
- Variation in allele frequencies in human subpopulations can mislead these analyses
- A disease with higher prevalence in one subpopulation than the other will look spuriously associated with alleles that have higher prevalence in the first
- Attenuation of a genetic effect is also possible - e.g., high prevalence in a small subpopulation

Impact of Structure on Association

Genetics Literature

- Considerable debate in the genetics community about the potential for spurious associations caused by population allele frequency differences (Wacholder et.al. 2000; Cardon and Palmer 2003; Freedmen et.al. 2004)
- Bidirectional effect - Structure can also mask or reverse genetic effects (Deng 2001)

Examples

- Knowler et.al. (1988) - Type 2 diabetes in American Indians
- Campbell et.al. (2005) - Lactase (LCT) gene and tall/short status in European Americans

Remedies

- Transmission Disequilibrium Test - Spielman et al. (1993)
- Genomic Control - Devlin & Roeder (1999)
- Modeled Classification - Pritchard et al. (2000)
- Structural Association - Satten et al. (2001)

Modeled Classification

Two-Stage Approach of Pritchard *et.al.*

- Assume we have a sample of N individuals, consisting of n_1 cases and n_0 controls
- Each subject is genotyped at L null marker loci and the candidate locus
- Assume each subject originates from one of K distinct subpopulations, where K is fixed.

Notation

- Population pointer
  $z_i \quad i = 1, ..., N$
- Genotype
  $X_{il}^{(c)} \quad l = 1, ..., (L+1)$
- Genotype frequencies
  $p_{klj} \quad j = 1, 2, 3$
- Admixture proportions $q_k^{(i)}$
- **Z**
- **X**
- **P**
- **Q**

Modeled Classification

STRUCTURE

- Intermediate Goal: Inference on **P** and **Z/Q** based on the genotypic data **X**
- Pritchard et al construct a Markov Chain Monte Carlo (MCMC) algorithm to approximate the posterior distribution

$$Pr\left(P, Z/Q|X\right) \propto Pr\left(Z/Q\right) Pr\left(P\right) Pr\left(X|P, Z/Q\right)$$

- Given **Z** and **P**, Pritchard et al then test
  - H_0 : No association between candidate allele frequencies and phenotype ϕ
  - H_1 : Candidate locus allele frequencies depend on phenotype ϕ

Modeled Classification

Likelihood Ratio Statistic: STRAT

- Define

$$\Lambda = \frac{Pr_1\left(X; \hat{P}_1, Q\right)}{Pr_0\left(X; \hat{P}_0, Q\right)}$$

- where

$$Pr_0^{(i)}\left[X_{i1}^{(c)} = j | P_0, \phi\right] = \sum_k q_k^{(i)} p_{k1j}$$

$$Pr_1^{(i)}\left[X_{i1}^{(c)} = j | P_1, \phi\right] = \sum_k q_k^{(i)} p_{k1j}^{\left[\phi^{(i)}\right]}$$

- The maximized likelihoods under each hypothesis are estimated using the EM algorithm
- Significance levels for Λ are approximated through simulation

Modeled Classification

Extensions and Other Approaches

- Falush *et.al.* (2003) - Extended Pritchard *et.al.*'s original work to allow for linkage disequilibrium and correlated allele frequencies
- Zhang *et.al.* (2006) - Parametric Bayesian approach that incorporates the inherent variability in estimating K through Model Averaging
- Other methods include Frequentist approaches such as Satten *et.al.* (2001), Epstein *et.al.* (2007), and Price *et.al* (2006) and a Bayesian approach for quantitative traits by Hoggart *et.al.* (2003)

Model : Quantitative Trait Part

Given a sample of N individuals, genotyped at L null marker loci and C candidates,

$$\begin{aligned} Y_i &\sim \text{Normal } (\mu_i, \tau_P) \\ \mu_i &= \beta_{0i} + G_i + \gamma Z_i \\ G_i &= \sum_{l=1}^{C} (\beta_{il} W_{il} + \beta_{2l} V_{il}) \end{aligned}$$

V and W are indicators of AA and aA respectively

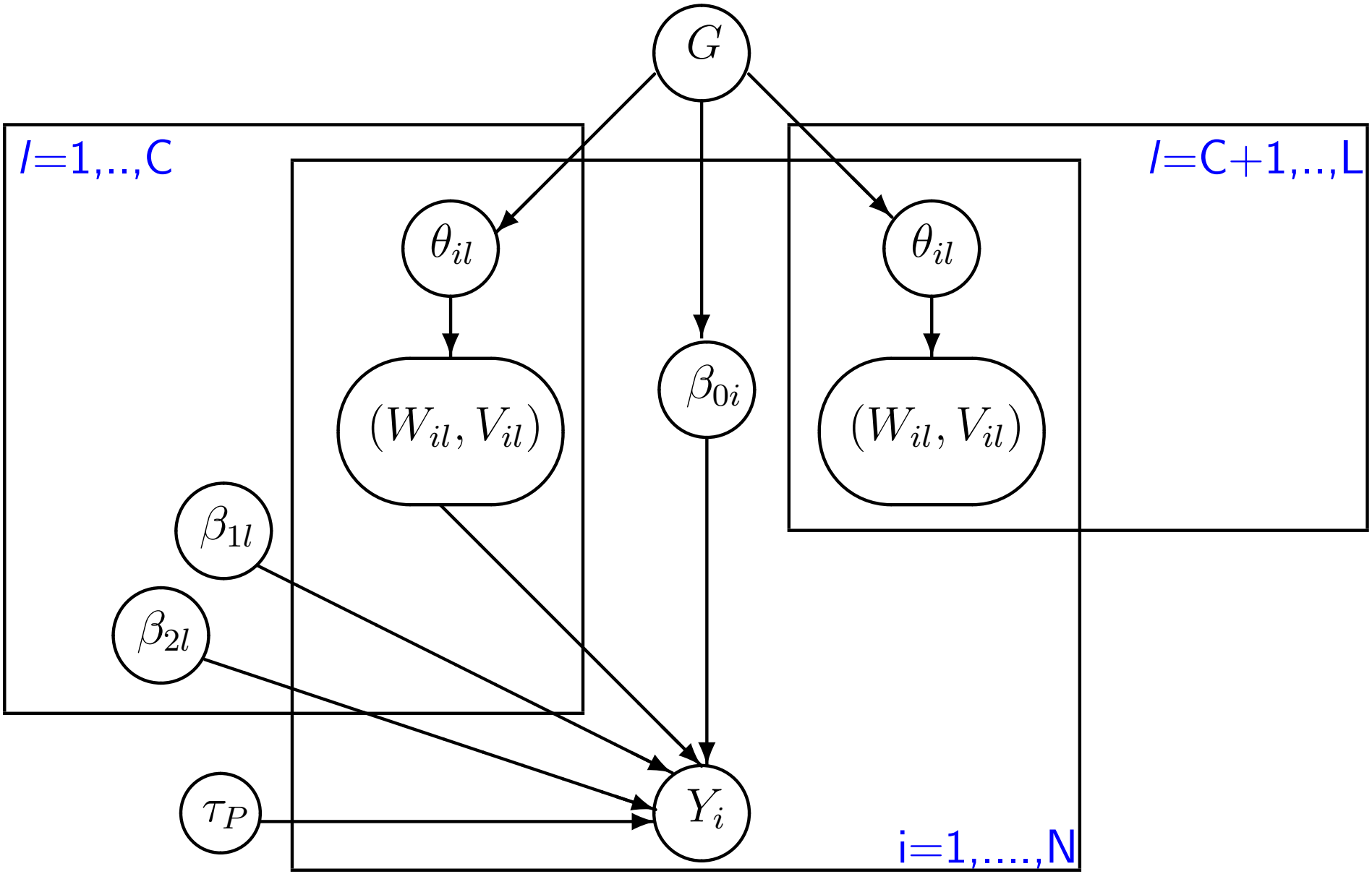
Model : Genetic Marker Part

Let $\theta_{il} = \text{logit}(\pi_{il})$,

$$L(W_i, V_i) = \prod_{l=1}^{L+C} \left[\frac{2^{W_{il}} \exp(\theta_{il}(2V_{il} + W_{il}))}{(1 + \exp(\theta_{il}))^2} \right]$$

$$\begin{aligned}
\beta_{0i}, \theta_{i1}, ..., \theta_{i,L+C} | G &\sim G \\
G | \alpha, G_0 &\sim \text{Dirichlet Process } (\alpha G_0) \\
\alpha &\sim \text{Gamma } (\alpha_0, \lambda_0) \\
G_0 &= f(\beta_0) \prod_{l=1}^{L+C} f(\theta_{il})
\end{aligned}$$

Model Graph


G
l=1,..,C
l=C+1,..,L
θ_{il}
θ_{il}
(W_{il}, V_{il})
β_{0i}
(W_{il}, V_{il})
β_{1l}
β_{2l}
τ_P
Y_i
i=1,....,N

Simulation of Posterior

Methods for DPM

Conjugate Setting

- Escobar (1994), Escobar and West (1995), Bush and MacEachern (1996)
- Split-Merge Samplers: Jain and Neal (2004), Green and Richardson (2001)

Non-conjugate Setting

- "No-Gaps" algorithm: MacEachern and Muller (1998)
- Neal (2000), Walker and Damien (1998)
- Jain and Neal (2007) extended their Split-Merge algorithm to a conditionally conjugate setting

Simulation of Posterior

Algorithm 5 of Neal (2000)

Given K current clusters of distinct values

1. Update configuration indicator for each individual (s_i) by proposing a new cluster membership (s_i^*) **R** times

$$\begin{aligned} P(s_i^* = j | s_{-i}) &= n_{-i,j}/(N + \alpha - 1) \text{ for j=1,..,K} \\ P(s_i^* = K+1 | s_{-i}) &= \alpha/(N + \alpha - 1) \end{aligned}$$

2. Use Metropolis-Hastings update for proposed move

$$a(s_i, s_i^*) = \min \left[\frac{L(Y_i, W_i, V_i | \phi_{s_i^*})}{L(Y_i, W_i, V_i | \phi_{s_i})}, 1 \right] \quad (1)$$

3. Update other parameters given current configuration. Combination of conjugate updates and Adaptive Rejection Sampling (ARS) (Gilks and Wild 1992)

Simulation of Posterior

MCMC Scheme

Given j=1,2,..,K distinct clusters in the $\mathcal{DP}$, let $\phi_j = (\beta_{0j}, \theta_{j1}, ..., \theta_{jL})$,

- Update configuration using Neal's Algorithm, sampling from $[s_i|(s_{-i}, \phi)]$ for i=1,....,N

For j=1,2,..,K,

- Sample $[\beta_{0j}|\mathbf{X}_j, \beta_{11}, .., \beta_{1C}, \beta_{21}, .., \beta_{2C}, \tau_P]$ where $\mathbf{X}_j = \{Y_i, W_{i\cdot}, V_{i\cdot}|s_i = j\}$ (conjugate normal update)
- For l=1,...,L, update cluster specific allele frequencies, sampling from $[\theta_{jl}|\mathbf{W}_j, \mathbf{V}_j]$ where $(\mathbf{W}_j, \mathbf{V}_j)$ denote genotype data for subjects in j^{th} cluster (ARS)

Simulation of Posterior

MCMC Scheme (continued)

Update genetic effects at candidate loci, for l=1,2,..,C

- Sample $\left[\beta_{1l}|\mathbf{Y},\mathbf{W},\beta_{01},..,\beta_{0K},\beta_{(1-l)},\beta_{21},..,\beta_{2C},\tau_P\right]$
- Sample $\left[\beta_{2l}|\mathbf{Y},\mathbf{V},\beta_{01},..,\beta_{0K},\beta_{(2-l)},\beta_{11},..,\beta_{1C}\right]$
  where $\beta_{(r-l)}=\beta_{(r1)},..,\beta_{(r,l-1)},\beta_{(r,l+1)},..,\beta_{(rC)}$
  (both are conjugate normal updates)

Update phenotypic precision

Sample $\left[\tau_P|\mathbf{Y},\mathbf{W},\mathbf{V},\beta_{01},..,\beta_{0K},\beta_{11},..,\beta_{1C},\beta_{21},..,\beta_{2C}\right]$
(conjugate gamma update)

A Two-Population Scenario

- Sample Size: 400 subjects with 500 markers
- Pop. 1 (50%) with $F_{ST} = 0.01$
- Pop. 2 (50%) with $F_{ST} = 0.035$

Simulated data characteristics
The biggest allele frequency difference,$\delta_l = |\pi_{1l} - \pi_{2l}|$, occurs at locus 254, where

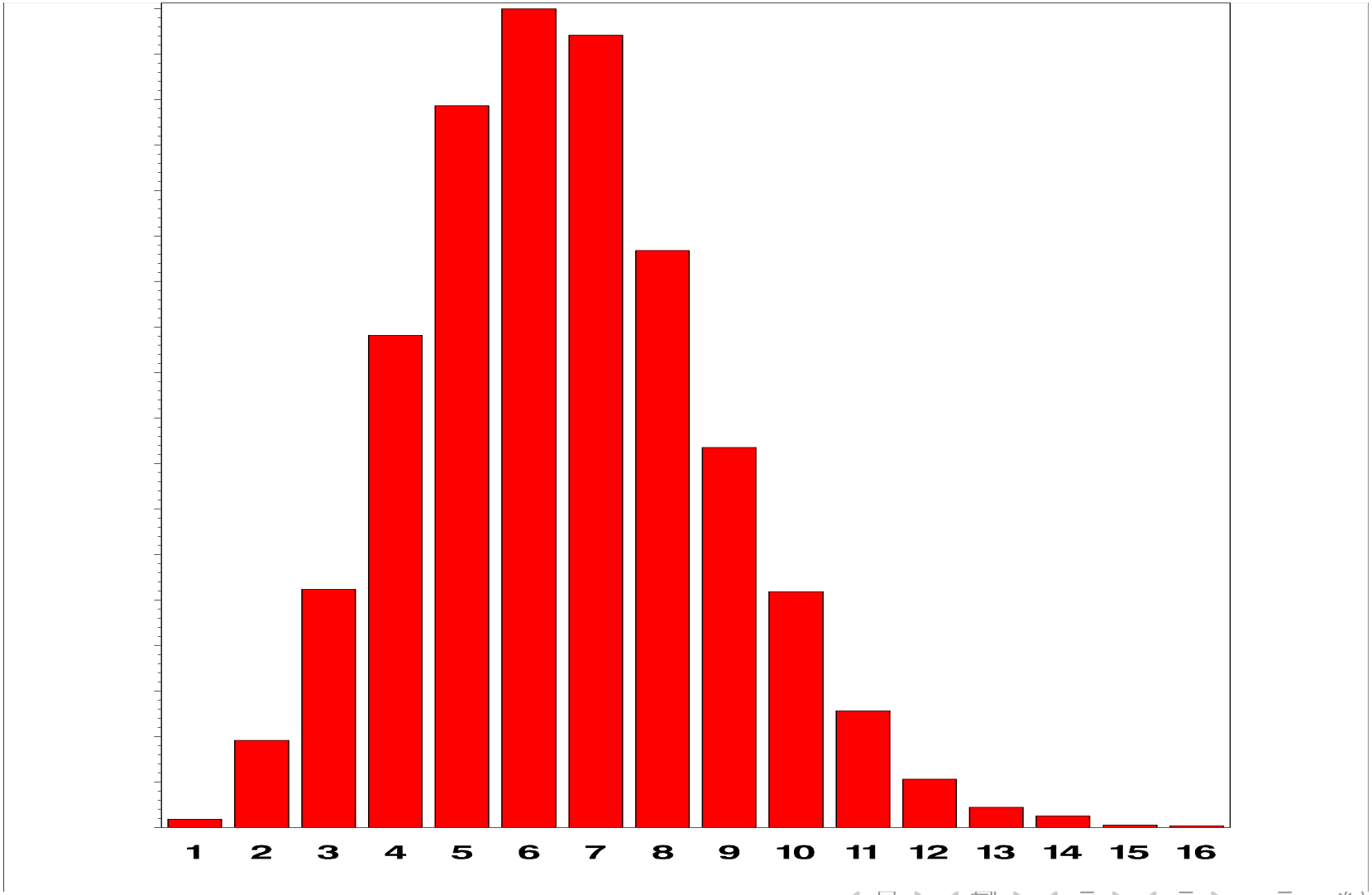
$$\delta_{254} = |0.713 - 0.358| = .355$$

The smallest difference occurs at locus 223,

$$\delta_{223} = |0.820 - 0.820| = .000$$

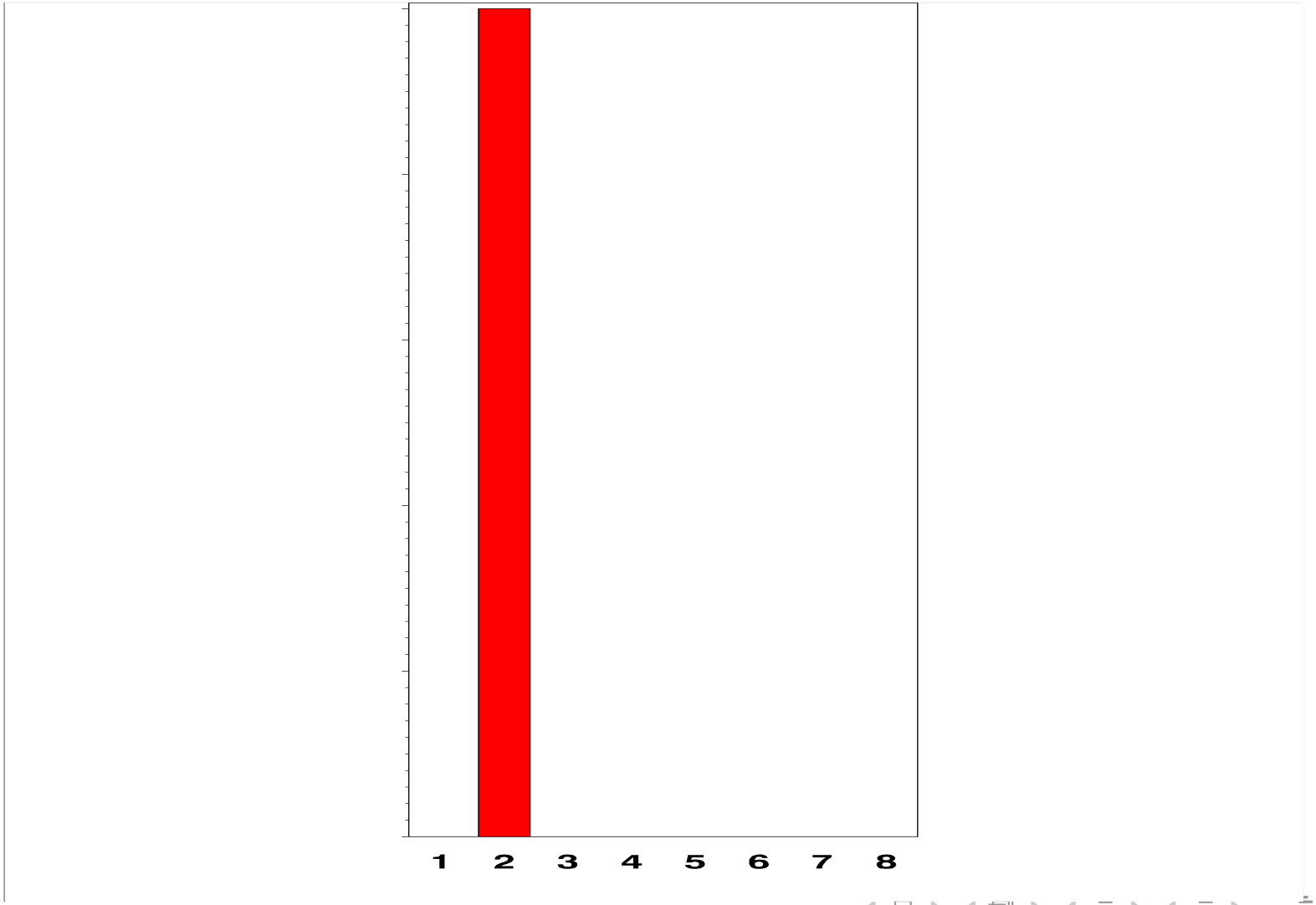
Prior for K

n=400 and α=1



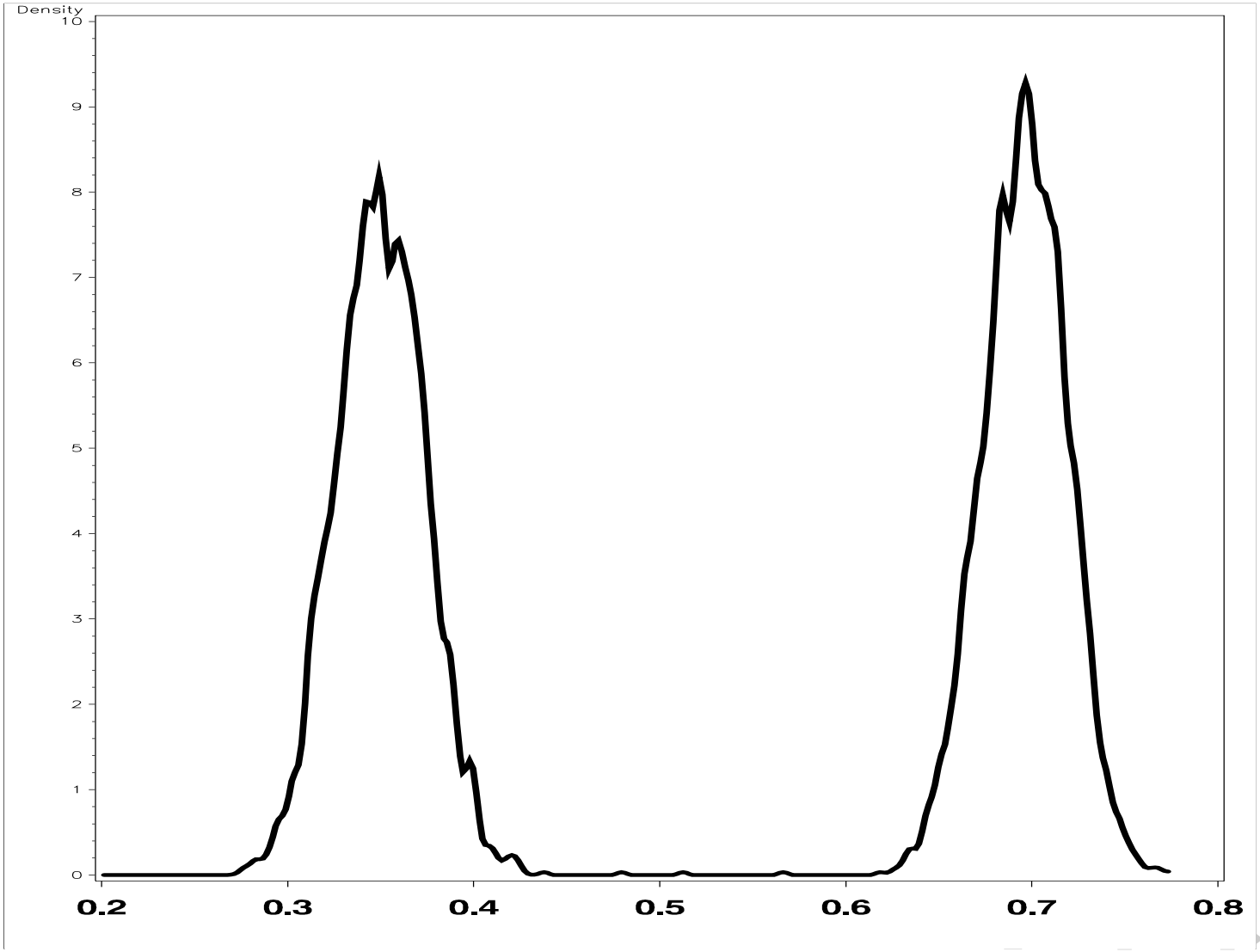
Posterior Distribution for K

Two populations scenario: n=400, L=500 and α=1



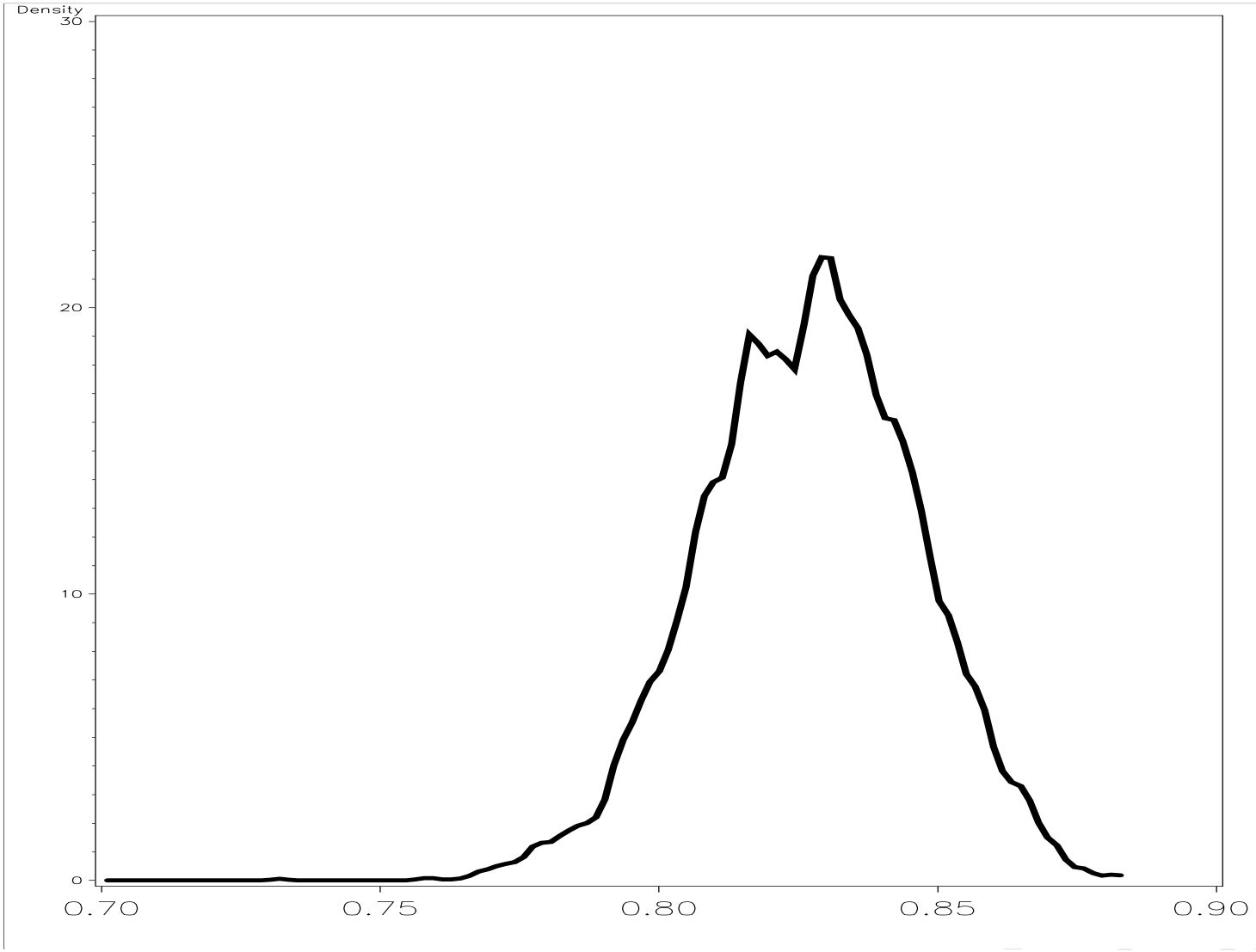
Predictive Density for π_{254}

$$\pi_{254}^{POP1} = 0.713 \text{ and } \pi_{254}^{POP2} = 0.358$$



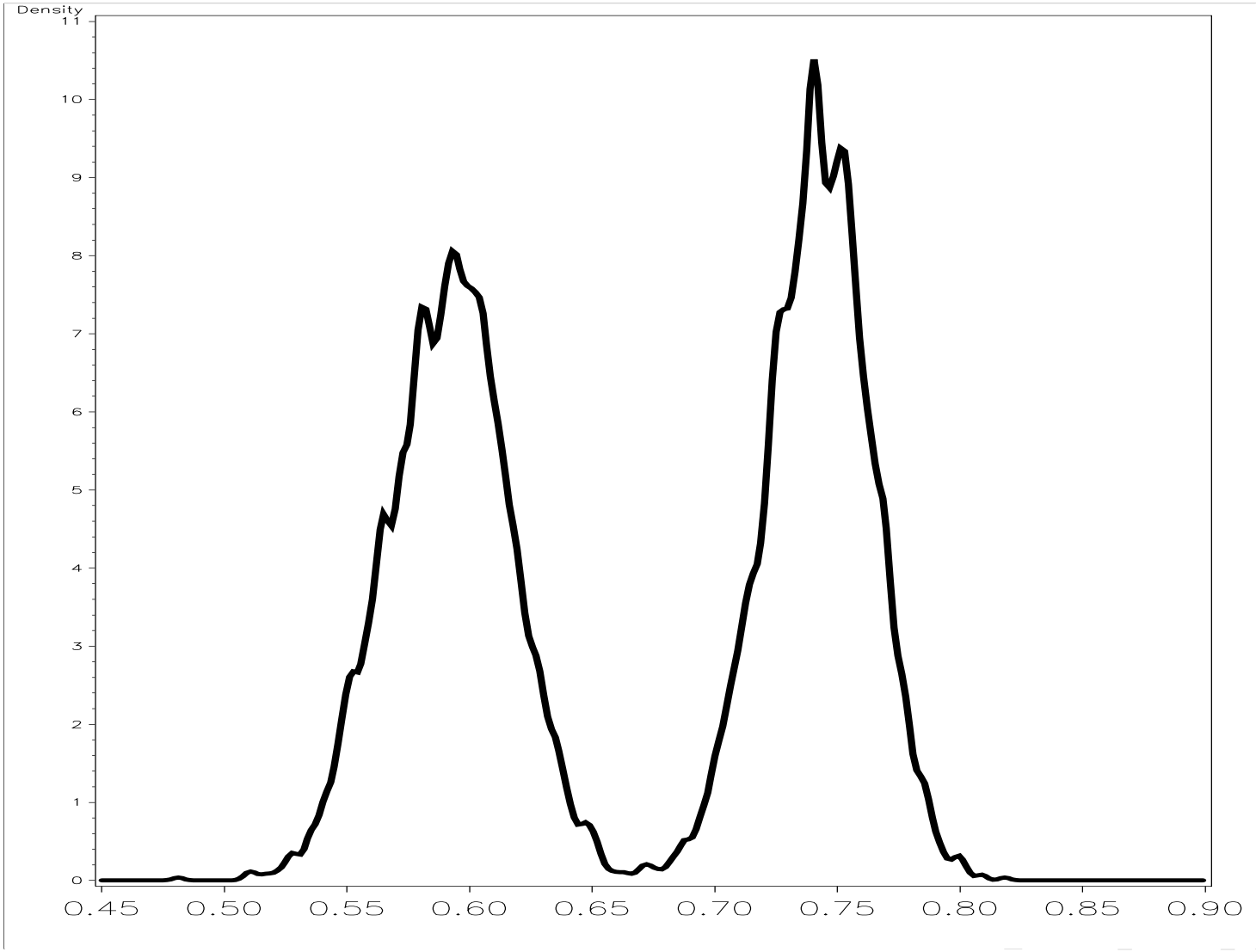
Predictive Density for π_{223}

$$\pi_{223}^{POP1} = 0.820 \text{ and } \pi_{223}^{POP2} = 0.820$$



Predictive Density for π_{109}

$\pi_{109}^{POP1} = 0.738$ and $\pi_{109}^{POP2} = 0.598$



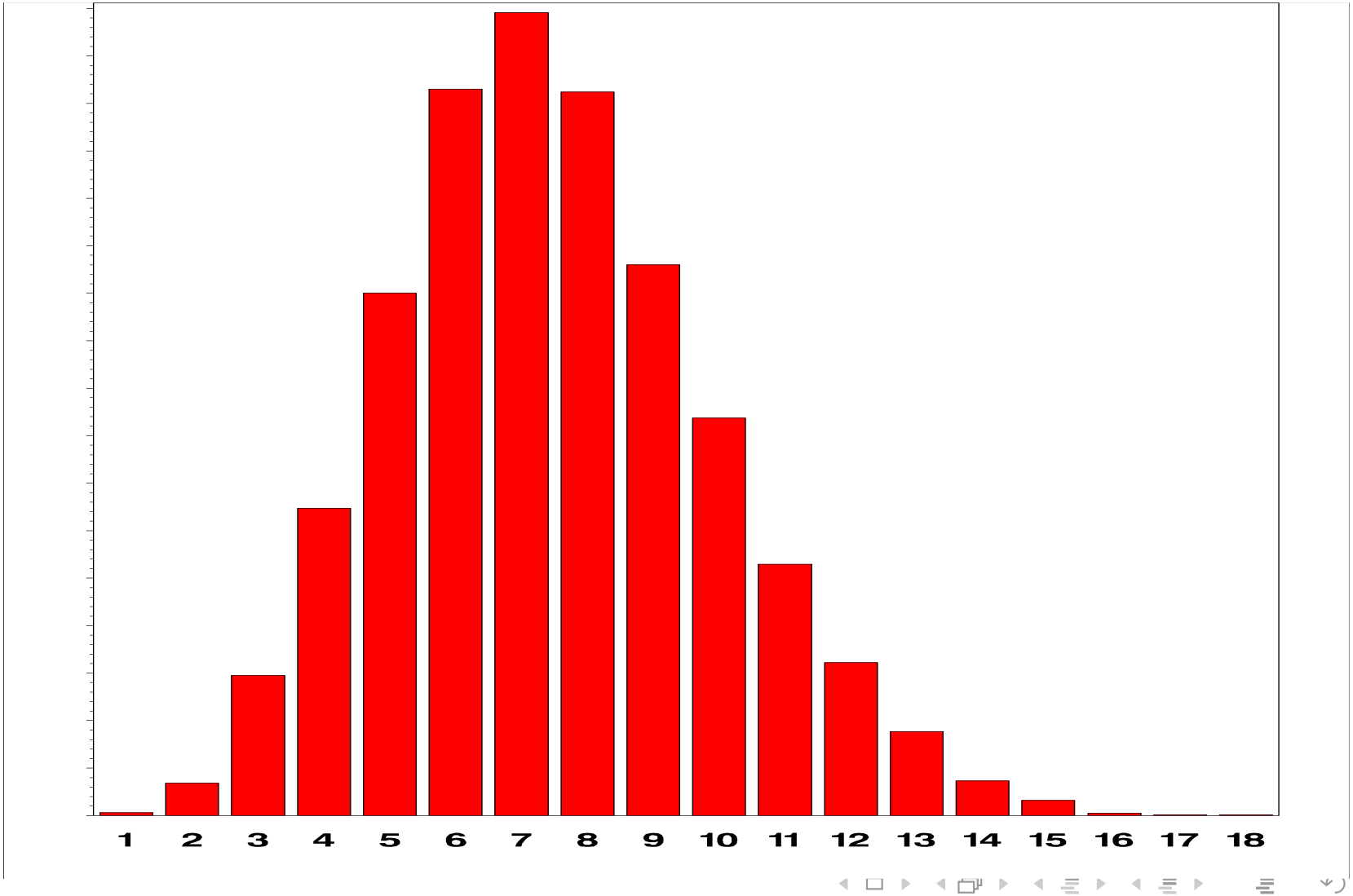
A Four-Population High Heterogeneity Scenario

Sample Size: 800 subjects at 1000 markers

Population	% of sample	$F_{ST}^{(j)}$
1	20	0.01
2	40	0.025
3	30	0.04
4	10	0.04

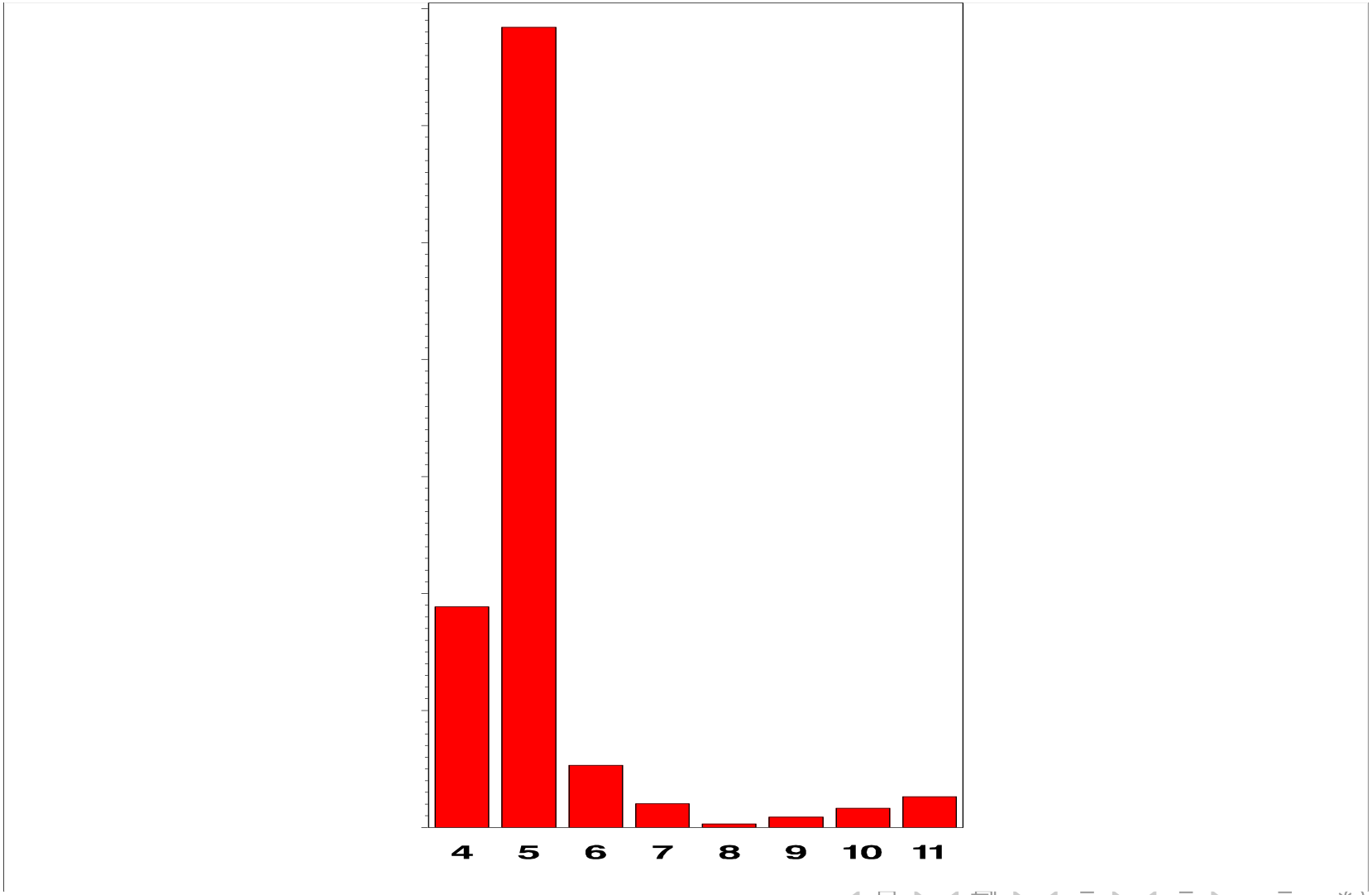
Prior for K

n=800 and α=1



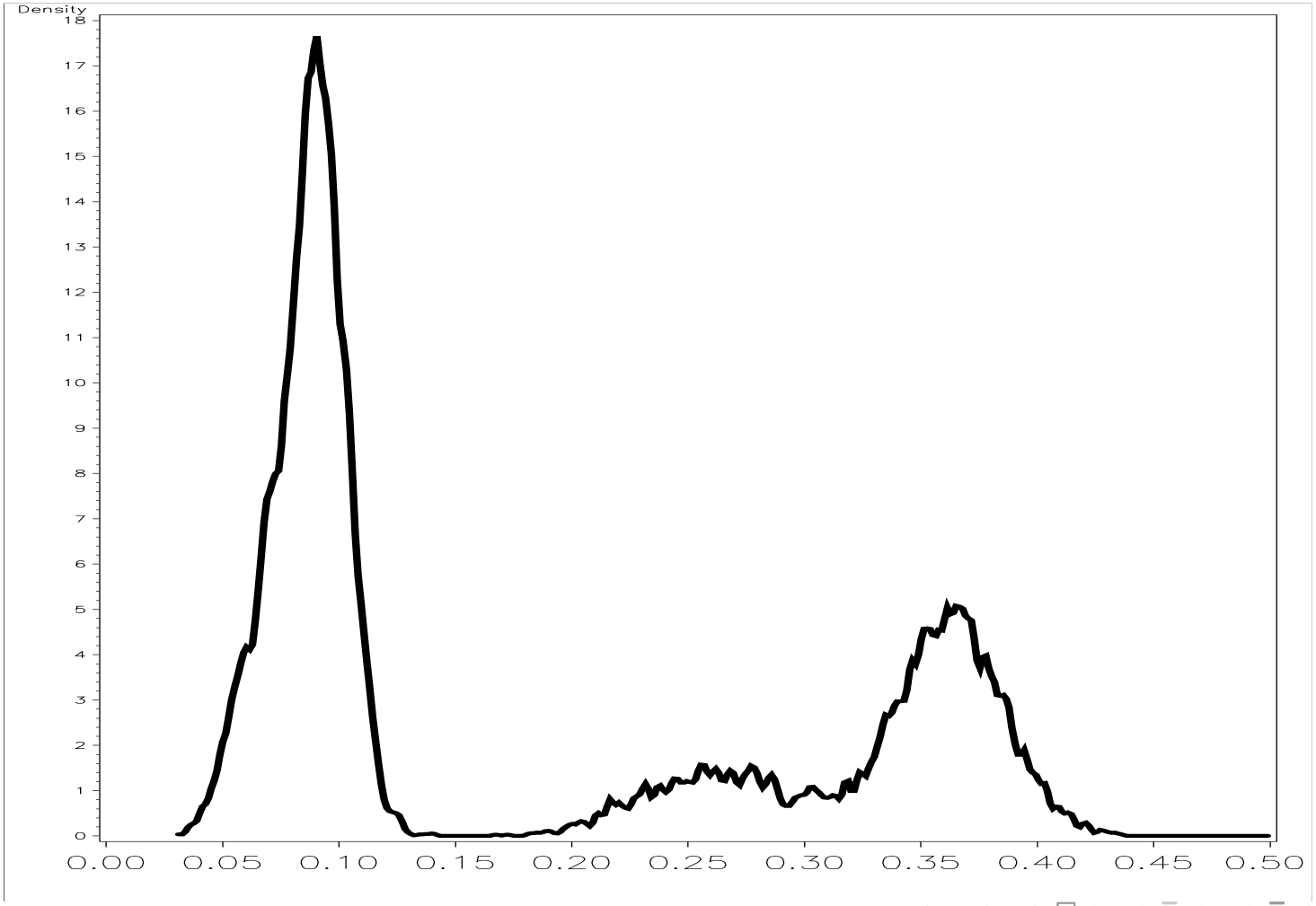
Posterior for K

4 Populations - High Heterogeneity



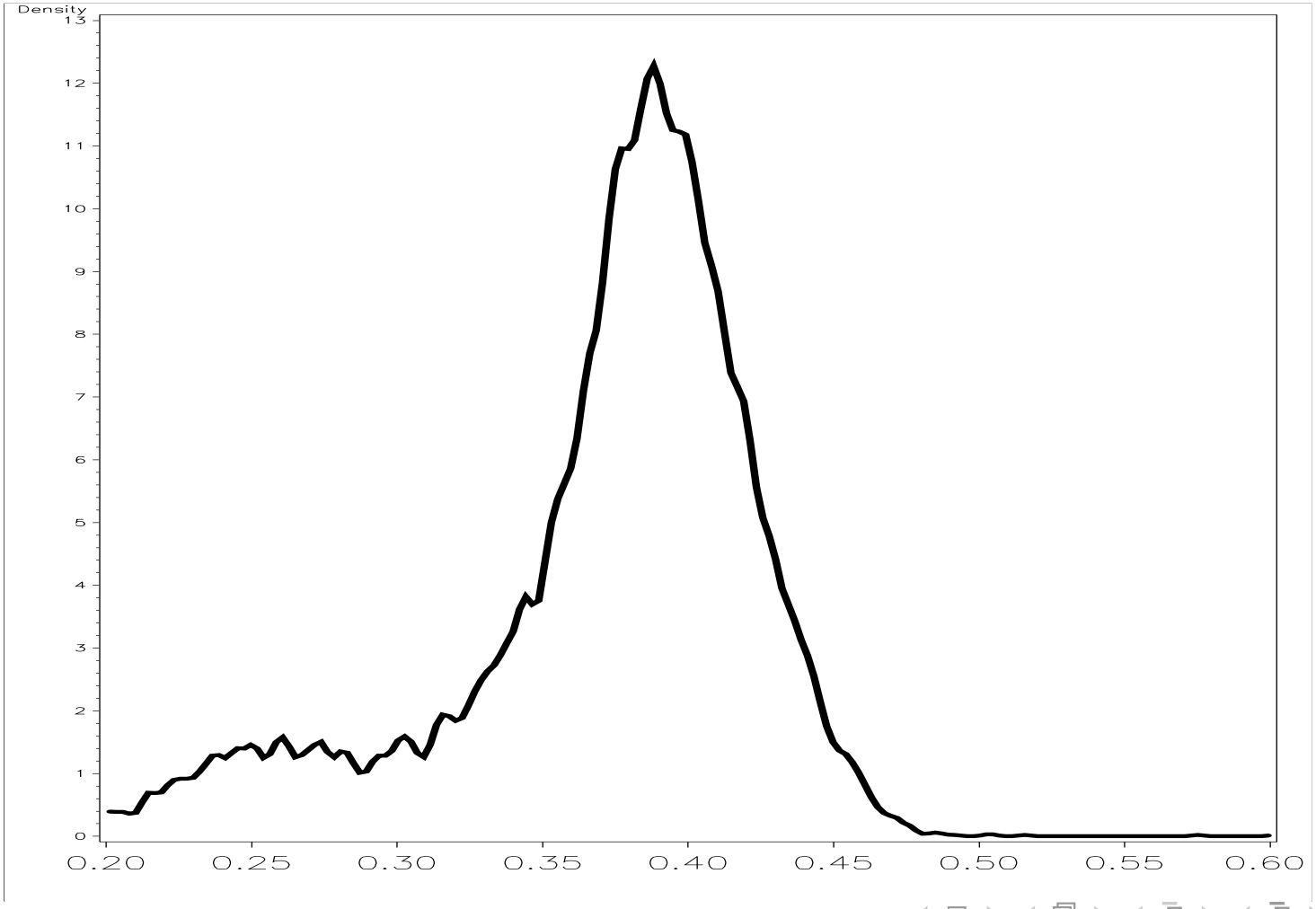
Predictive Density for π_{43}

$$(20\%)\ \pi^{(1)} = 0.07\ (40\%)\ \pi^{(2)} = 0.08$$
$$(30\%)\ \pi^{(3)} = 0.37\ (10\%)\ \pi^{(4)} = 0.24$$



Predictive Density for π_{47}

$$(20\%)\ \pi^{(1)} = 0.34\ (40\%)\ \pi^{(2)} = 0.39$$
$$(30\%)\ \pi^{(3)} = 0.42\ (10\%)\ \pi^{(4)} = 0.34$$



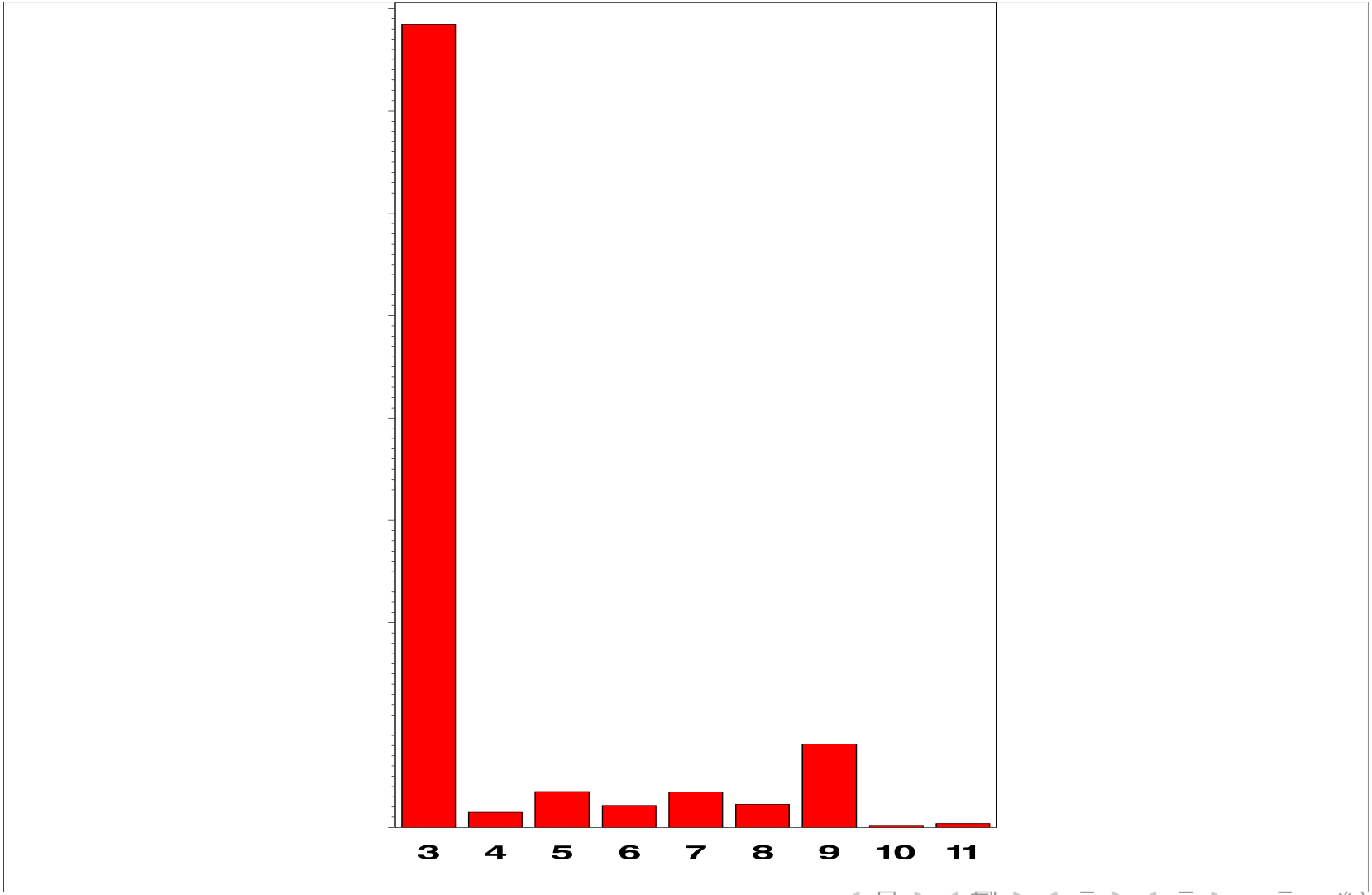
A Four-Population Low Heterogeneity Scenario

Sample Size: 800 subjects at 1000 markers

Population	% of sample	$F_{ST}^{(j)}$
1	40	0.01
2	10	0.02
3	20	0.04
4	30	0.04

Posterior for K

4 Populations - Low Heterogeneity



Quantitative Trait Model Simulation

- Three scenarios where individuals originate from two ancestral populations
- Baseline trait mean differs between populations, with $\beta_{01} = 100.0$ and $\beta_{02} = 120.0$
- Subjects genotyped at 750 markers including 5 candidates for association with the phenotype (generically denoted as loci 1-5)

Scenarios

- I. n=700; Pop. 1 (50%) and Pop. 2 (50%)
- II. n=1000; Pop. 1 (50%) and Pop. 2 (50%)
- III. Same as I plus 300 admixed individuals

Simulation Setup

Locus	β_{1l}	β_{2l}	A_l	D_l
1	0	10	5	-5
2	0	5	2.5	-2.5
3	0	0	0	0
4	7.5	7.5	3.75	3.75
5	0	0	0	0

$$
\begin{aligned}
Y_i &\sim \text{Normal } (\mu_i, \tau_P) \text{ where} \\
\mu_i &= \beta_{0i} + \sum_{l=1}^{5} (\beta_{1l} W_{il} + \beta_{2l} V_{il}) \\
\tau_P &= 0.01
\end{aligned}
$$

A_l and D_l = Additive / Dominance effects

Falconer & MacKay (1996)

Generation of Admixed Ancestry Individuals

- Assume j=1,..,K subpopulations with associated allele frequencies π_{jl} at each locus
- Let $q_{i\cdot}$ denote an admixture vector where q_{ij} denotes the proportion of the i^{th} admixed individual's genome that originates from the j^{th} subpopulation.
- $q_{i\cdot} \sim$ Dirichlet$(\alpha_q, ..., \alpha_q)$ where α_q was set at 1.0
- Then the genotype at the l^{th} locus is independently simulated as Binomial$(2, \sum_{j=1}^{K} q_{ij}\pi_{jl})$

Naive Analysis

Effect	True	$\hat{\beta}$	95% CI*	$\chi^2_{2,0.99} = 9.28$
β_{11}	0	-4.08	-11.88 to 3.72	43.95
β_{21}	10	2.67	-4.89 to 10.22	
β_{12}	0	-7.23	-9.93 to -4.53	67.51
β_{22}	5	-8.29	-13.13 to -3.46	
β_{13}	0	-1.40	-4.51 to 1.70	1.71
β_{23}	0	-1.27	-4.77 to 2.24	
β_{14}	7.5	9.25	2.03 to 16.48	13.69
β_{24}	7.5	7.41	0.44 to 14.38	
β_{15}	0	5.83	-1.87 to 13.53	5.68
β_{25}	0	6.28	-1.14 to 13.69	

*Simultaneous CI using Bonferroni Correction

Scenario I Results

Two Distinct Subpopulations (n=700)

Parameter	True	$\hat{\beta}$	95% Cred. Int.*	χ^2_l
β_{11}	0	1.06	-4.54 to 6.63	131.37
β_{21}	10	10.86	5.26 to 16.63	
β_{12}	0	1.25	-1.41 to 3.81	4.26
β_{22}	5	3.06	-0.99 to 7.00	
β_{13}	0	-0.60	-3.21 to 1.74	1.43
β_{23}	0	0.46	-3.34 to 3.33	
β_{14}	7.5	9.24	3.22 to 13.80	20.10
β_{24}	7.5	9.09	3.23 to 13.25	
β_{15}	0	-0.32	-6.61 to 5.88	15.35
β_{25}	0	-3.80	-9.14 to 1.93	

* denotes simultaneous 95% Credible Intervals

Scenario II Results

Two Distinct Subpopulations (n=1000)

Parameter	True	$\hat{\beta}$	95% Cred. Int.*	χ^2_l
β_{11}	0	0.31	-4.68 to 5.27	199.57
β_{21}	10	10.42	6.12 to 15.26	
β_{12}	0	0.56	-1.34 to 2.68	7.64
β_{22}	5	3.71	-0.13 to 7.35	
β_{13}	0	0.27	-2.32 to 2.55	2.26
β_{23}	0	1.31	-1.68 to 4.13	
β_{14}	7.5	7.31	3.47 to 12.28	21.37
β_{24}	7.5	7.60	3.93 to 12.41	
β_{15}	0	1.81	-2.29 to 6.41	6.48
β_{25}	0	-0.16	-4.10 to 4.64	

* denotes simultaneous 95% Credible Intervals

Scenario III Results

n=1000, Including 300 Admixed Individuals

Parameter	True	$\hat{\beta}$	95% Cred. Int.*	χ^2_l
β_{11}	0	0.64	-7.17 to 6.82	58.52
β_{21}	10	7.14	-0.01 to 13.00	
β_{12}	0	0.08	-2.31 to 2.51	3.11
β_{22}	5	2.83	-1.79 to 7.06	
β_{13}	0	-1.86	-4.21 to 0.74	3.79
β_{23}	0	-0.81	-3.43 to 2.31	
β_{14}	7.5	6.54	-0.52 to 12.39	5.89
β_{24}	7.5	5.94	-0.30 to 11.42	
β_{15}	0	1.80	-4.79 to 10.68	7.33
β_{25}	0	-0.83	-6.90 to 7.48	

* denotes simultaneous 95% Credible Intervals

Comparative Results

Parameter	True	N	S I	S II	S III
β_{11}	0	-4.08	1.06	0.31	0.64
β_{21}	10	2.67	10.86	10.42	7.14
β_{12}	0	-7.23	1.25	0.56	0.08
β_{22}	5	-8.29	3.06	3.71	2.83
β_{13}	0	-1.40	-0.60	0.27	-1.86
β_{23}	0	-1.27	0.46	1.31	-0.81
β_{14}	7.5	9.25	9.24	7.31	6.54
β_{24}	7.5	7.41	9.09	7.60	5.94
β_{15}	0	5.83	-0.32	1.81	1.80
β_{25}	0	6.28	-3.80	-0.16	-0.83

Model for Admixture

- Inclusion of highly admixed individuals affects estimation
- The DPM construction above clusters individuals based on their entire set of genotypes at all L loci. It does not allow for differing ancestries across the genome
- How to handle admixed individuals more appropriately?
- Assume, for a moment, that we know which individuals are highly admixed. Divide the sample between N_L low admixture subjects and N_H high admixture subjects
- We then consider a separate parametric likelihood for the N_H individuals that uses locus-specific information, effectively making an assignment at each locus for each admixed person to the "clusters" detected by the Dirichlet Process

Admixture Model Likelihood

For the i^{th} admixed subject, let $X_{il}^{(c)}$ denote the allele at the l^{th} locus on the c^{th} chromosome, $c = 1, 2$. Let $z_{il}^{(c)}$ denote a subpopulation assignment indicator for that allele
Let q_{ij} represent an N_L dimensional admixture vector that denotes how much of that person's genome originates from the subpopulations represented by the low admixture sample.

$$\begin{aligned}\mathbb{L}_H &= \prod_{i=1}^{N_H}\left[\frac{\tau_P^{1/2}}{\sqrt{2\pi}}\exp\left(\frac{-\tau_P}{2}\left(Y_i - \sum_{j=1}^{N_L} q_{ij}\beta_{0j} - G_i\right)^2\right)\right] \\ &\times \prod_{l=1}^{L}\prod_{c=1}^{2}\prod_{j=1}^{N_L}\left[\left(\frac{q_{ij}e^{\theta_{jl}X_{il}^{(c)}}}{1+e^{\theta_{jl}}}\right)^{I(z_{il}^{(c)}=j)}\right]\end{aligned}$$

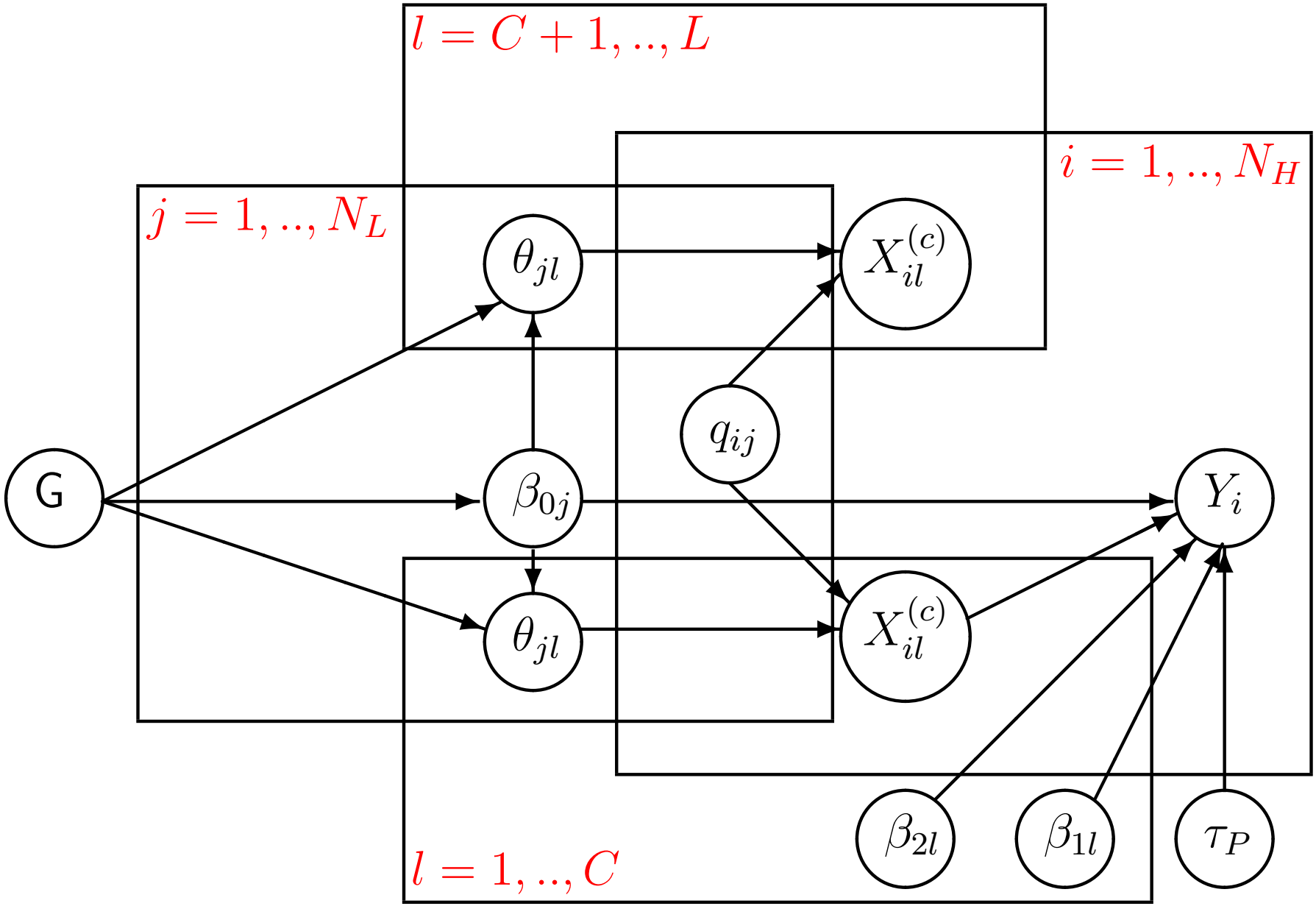
Admixture Model Continued

The model is then the same as before except for the addition of two prior specifications

- $q_{i\cdot} \sim$ Dirichlet $(\alpha_1, ..., \alpha_{N_H})$ for $i = 1, ..., N_H$
- Pr$(z_{il}^{(c)} = j) = \frac{1}{N_L}$ for $j = 1, ..., N_L$

Note: Although the model is setup to assign alleles to specific low-admixture individuals, because of the inherent Dirichlet Process clustering, the calculations simplify to assigning "admixed" alleles to the current K distinct clusters

Admixture Model Graph



Scenario III Results With Hybrid Model

Parameter	True	$\hat{\beta}$	95% Cred. Int.*	χ_l^2
β_{11}	0	1.36	-4.18 to 8.07	62.74
β_{21}	10	8.14	2.71 to 14.76	
β_{12}	0	0.48	-2.23 to 2.84	4.42
β_{22}	5	3.63	-1.12 to 8.17	
β_{13}	0	-1.28	-3.89 to 1.48	1.98
β_{23}	0	-0.34	-3.19 to 2.73	
β_{14}	7.5	7.52	-0.77 to 14.46	10.18
β_{24}	7.5	6.76	-1.66 to 12.99	
β_{15}	0	1.95	-6.38 to 9.22	7.40
β_{25}	0	-0.64	-8.76 to 6.33	

* denotes simultaneous 95% Credible Intervals

Comparative Results

Parameter	True	N	S I	S II	S III	S IIIA
β_{11}	0	-4.08	1.06	0.31	0.64	1.36
β_{21}	10	2.67	10.86	10.42	7.14	8.14
β_{12}	0	-7.23	1.25	0.56	0.08	0.48
β_{22}	5	-8.29	3.06	3.71	2.83	3.63
β_{13}	0	-1.40	-0.60	0.27	-1.86	-1.28
β_{23}	0	-1.27	0.46	1.31	-0.81	-0.34
β_{14}	7.5	9.25	9.24	7.31	6.54	7.52
β_{24}	7.5	7.41	9.09	7.60	5.94	6.76
β_{15}	0	5.83	-0.32	1.81	1.80	1.95
β_{25}	0	6.28	-3.80	-0.16	-0.83	-0.64

Detecting Admixed Individuals

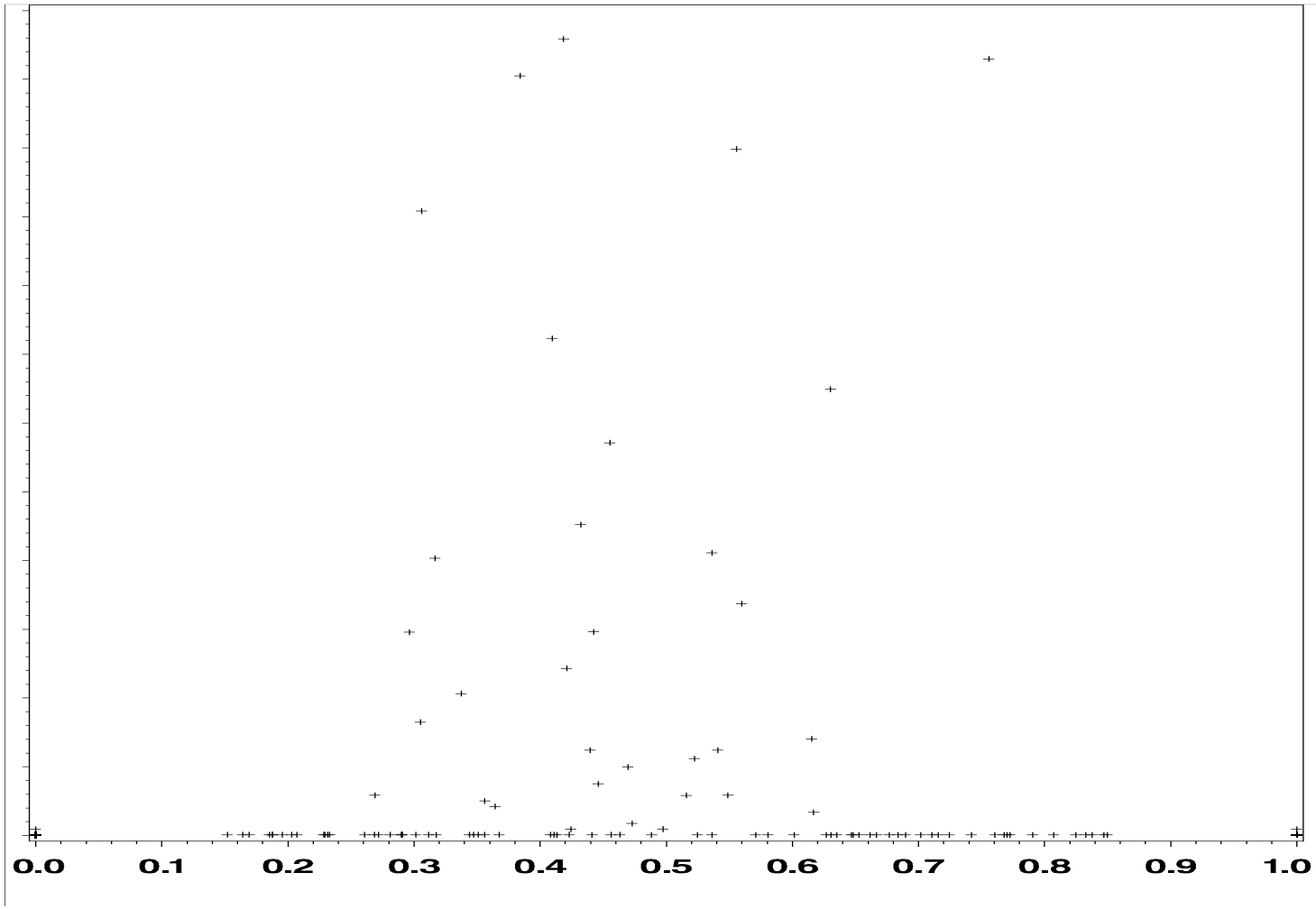
Let $s_i^{(g)}$ denote the cluster membership indicator for the i^{th} individual at the g^{th} MCMC iteration, then define

$$\hat{\eta_{il}}^{(g)} = \frac{\sum_{\left\{j:s_j^{(g)}=s_i^{(g)}\right\}}^{N} (2V_{jl} + W_{jl})}{2n_{s_i^{(g)}}}$$

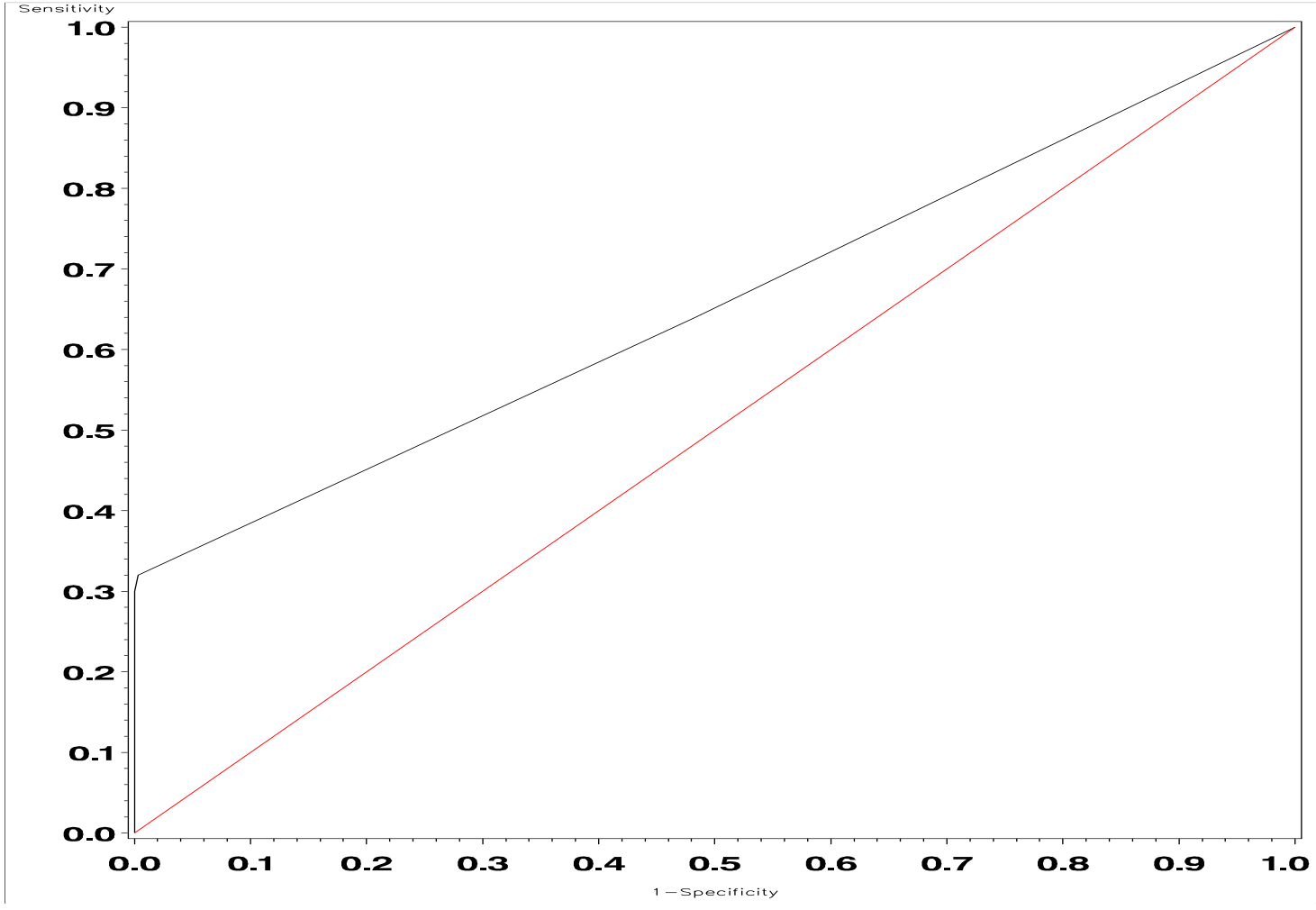
Then let

$$\epsilon_i^{(g+1)} = \sum_{l=1}^{L} \left(\hat{\eta}_{il}^{(g+1)} - \hat{\eta}_{il}^{(g)}\right)^2$$

Results on Detection



Results on Detection



Modifications for Case-Control Studies

Binary Outcomes

Simply involves switching to a logistic or probit likelihood.

Case-Control Studies

- The extension to a case-control setting is also relatively straightforward
- Need to specify a retrospective likelihood such as that proposed by
  - Seaman and Richardson (2001)
  - Müller and Roeder (1997)

Extensions

1. Clustering by genome sections

2. Introducing dependence structure amongst neighboring loci, i.e. incorporating linkage disequilibrium patterns

3. Extending to Genomewide Association Studies - Removing candidate loci designation so that all markers are potentially phenotypically associated

References

In a separate file