

Haplotype Estimation and AFLP Mixture Models

Paul Eilers^{1,2} & Hae-Won Uh³

¹*Erasmus Medical Center, Rotterdam, The Netherlands*

²*Biometris, Wageningen, The Netherlands*

³*Leiden University Medical Centre, Leiden, The Netherlands*

Haplotype and diplotype probabilities

- Consider one marker, alleles A and B
- Indicate haplotype (= one chromosome) by number of B alleles
- Only two possibilities, 0 or 1
- Probabilities p_0 and $p_1 = 1 - p_0$
- Two chromosomes, hence four ordered pairs, the diplotypes
- Assume random pairing of haplotypes
- The diplotype probabilities then are p_0^2 , p_0p_1 , p_1p_0 and p_1^2

More markers

- Two markers, four haplotypes: 00, 01, 10, and 11
- Let their probabilities be π_1 to π_4
- Introduce logs of probabilities: $\log \pi_k = \beta_k$
- Now we can write diplotype probabilities as $\gamma = \exp(X\beta)$

$$X' = \begin{pmatrix} 2 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 2 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 2 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 2 \end{pmatrix}$$

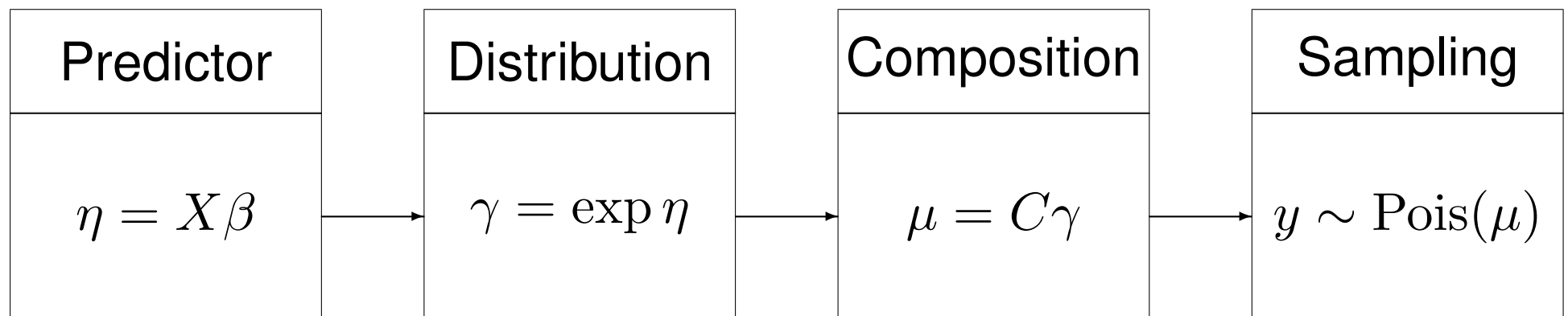
- This can be generalized to more markers

We cannot observe diplotypes

- We cannot discriminate order in pairs: $\pi_j\pi_k = \pi_k\pi_j$
- Instead we observe unordered pairs, the genotypes
- Introduce $C = [C_{ij}]$, the *composition* matrix
- $c_{ij} = 1$ if diplotype j gives genotype i , otherwise $c_{ij} = 0$
- Consider n individuals
- Expected numbers of genotypes: $\mu = nC\gamma$
- For L markers, C has 3^L rows and 2^{2L} columns

The composite link model

- We have here a composite link model (CLM)
- Proposed by Thompson and Baker (1981), but largely neglected
- Elegant framework for modeling indirect observations of counts



Maximum likelihood estimation

- Poisson log-likelihood: $\sum_i (y_i \log \mu_i - \mu_i)$
- Thompson and Baker gave algorithm
- Treat CLM as GLM with working design matrix $U = M^{-1}C\Gamma X$
- Where $M = \text{diag}(\mu)$ and $\Gamma = \text{diag}(\gamma)$
- This works fine for well-conditioned problems
- The haplotype problem is ill-conditioned
- Some β can approach $-\infty$

Penalized likelihood

- To stabilize the model we introduce a penalty
- Let α be some “sensible” vector (to be explained later)
- Write $\beta = \beta^* + \delta$, push β^* (more or less gently) towards α
- We need δ to make $\sum_k \exp \beta_k = 1$
- A penalty, $\kappa \sum_k (\beta_k^* - \alpha_k)^2 / 2$, achieves this
- The larger κ , the nearer β^* will be to α
- Penalized log-likelihood

$$\sum_i (y_i \log \mu_i - \mu_i) - \kappa \sum_k (\beta_k^* - \alpha_k)^2 / 2$$

A reasonable goal for shrinking

- The penalty shrinks β^* towards α
- What will we use for α ?
- Reasonable choice: assume all markers to be independent
- No linkage disequilibrium
- Haplotype probability then is product of allele probabilities
- $\prod_l \pi_l^{t_l} (1 - \pi_l)^{1-t_l}$ for haplotype $t_1 t_2 \dots t_L$

How much shrinkage?

- How do we choose κ ?
- AIC (Akaike's Information Criterion) is attractive
- $AIC = \text{Deviance} + 2 * ED$
- ED is effective model dimension
- We will see it below

The scoring algorithm

- We skip the technical details
- The scoring algorithm tells us to repeatedly solve

$$(\tilde{U}'\tilde{W}\tilde{U} + \kappa I)\hat{\beta} = \tilde{U}'(y - \tilde{\mu} + \tilde{W}\tilde{U}\tilde{\beta}) + \kappa\alpha,$$

- Where a tilde indicates current approximation, $U = M^{-1}C\Gamma X$
- And $M = \text{diag}(\mu)$, $\Gamma = \text{diag}(\gamma)$ and $W = \text{diag}(\mu)$.
- At convergence, $\text{cov}(\hat{\beta}) = (\hat{U}'\hat{W}\hat{U} + \kappa I)^{-1}$
- Effective model dimension: $\text{ED} = \text{trace}[(\hat{U}'\hat{W}\hat{U} + \kappa I)^{-1}\hat{U}'\hat{W}\hat{U}]$

Some practical details

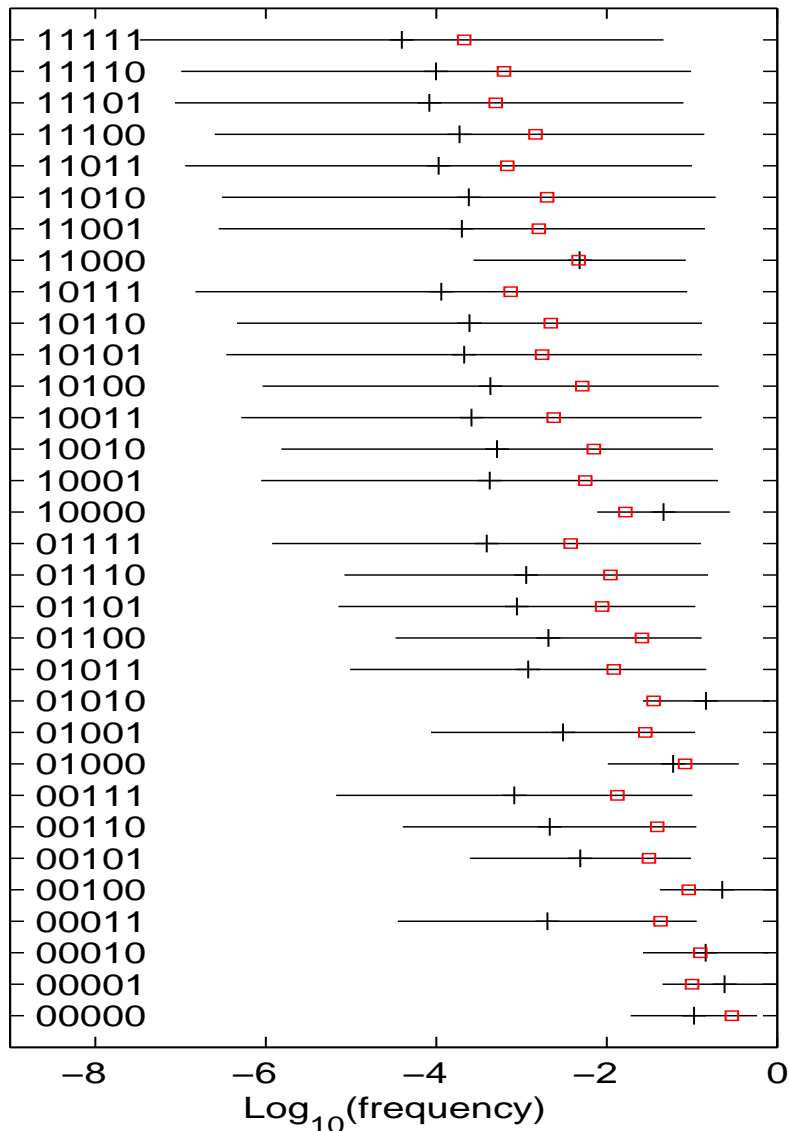
- The systems of equations are large
- With 10 markers we have 1024 haplotypes
- Over a million diplotypes and 59049 possible genotypes
- Both X and C are very sparse
- Matlab handles sparse matrices transparently and efficiently
- But U generally not very sparse
- So we have to solve a full system with 1024 unknowns
- Line search (step halving) necessary for stable scoring algorithm

An example from human genetics

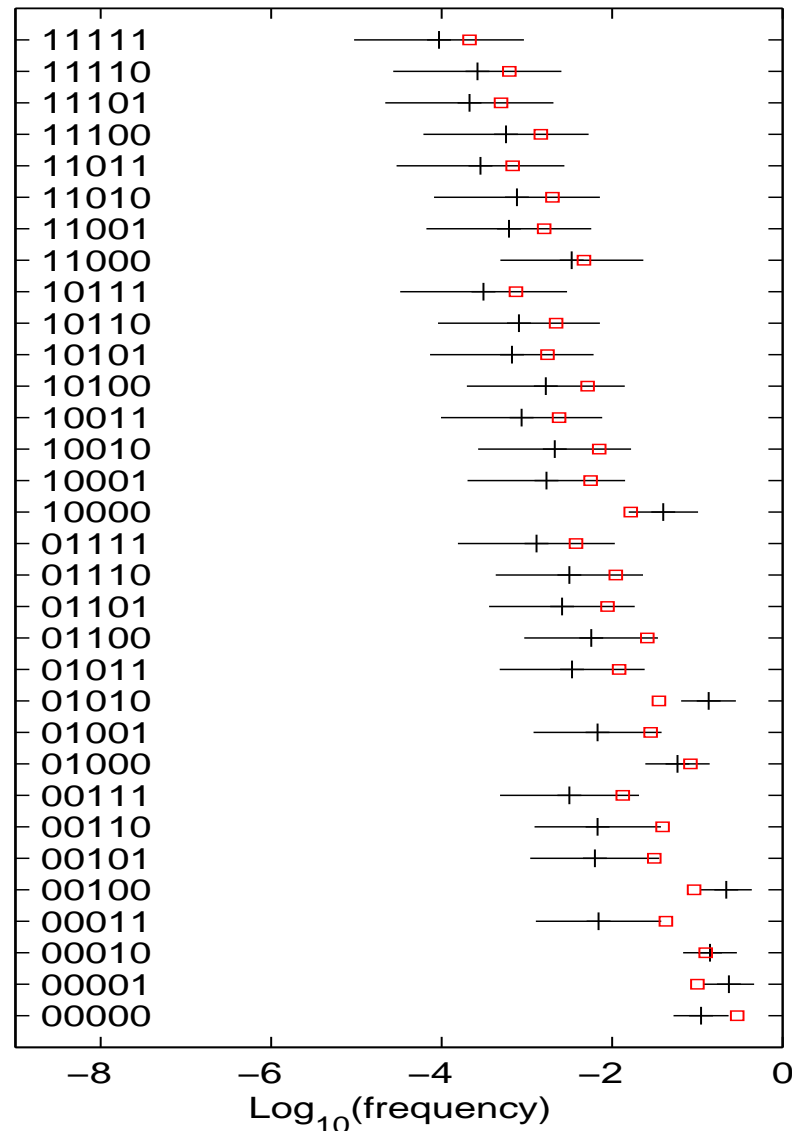
- Antigen Processing Machinery (APM) data
- Mehta et al., 2007
- Case-control study on cervical carcinoma
- 5 SNPs on chromosome 5
- 122 Controls, 86 cases

Estimated probabilities in controls

Estimates for $\kappa = 0.1$; CC = 0

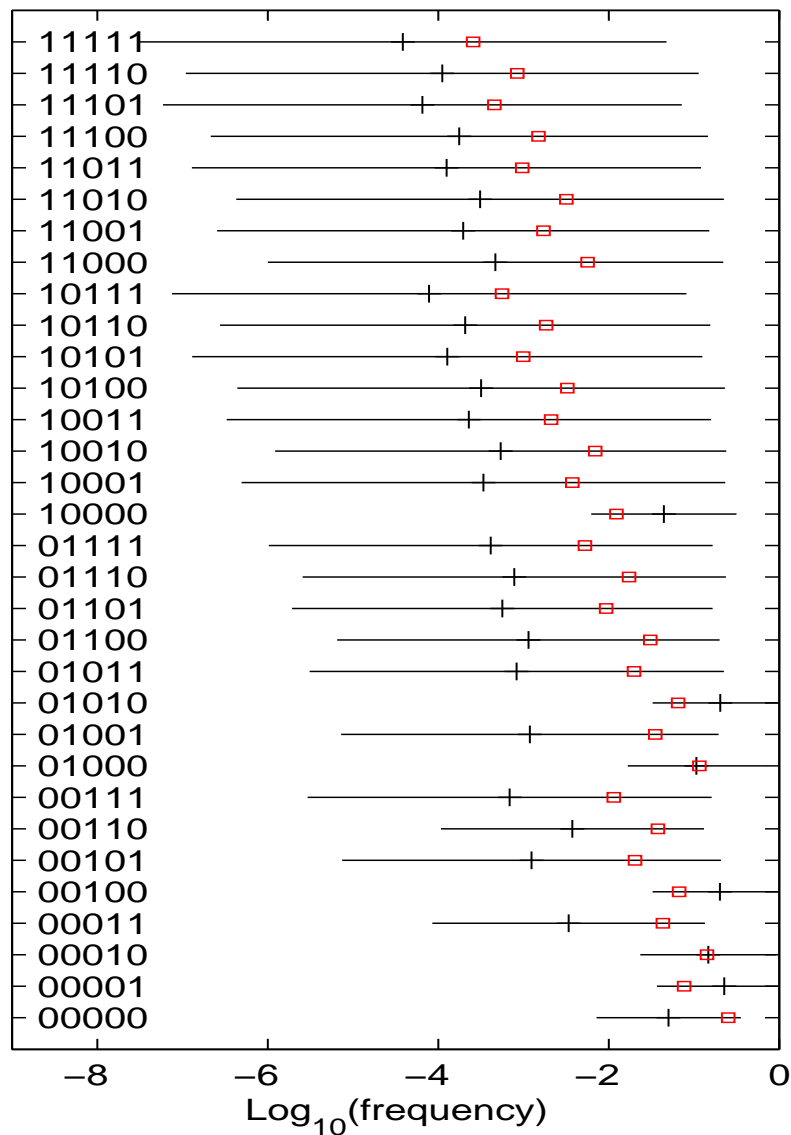


Estimates for $\kappa = 1$; CC = 0

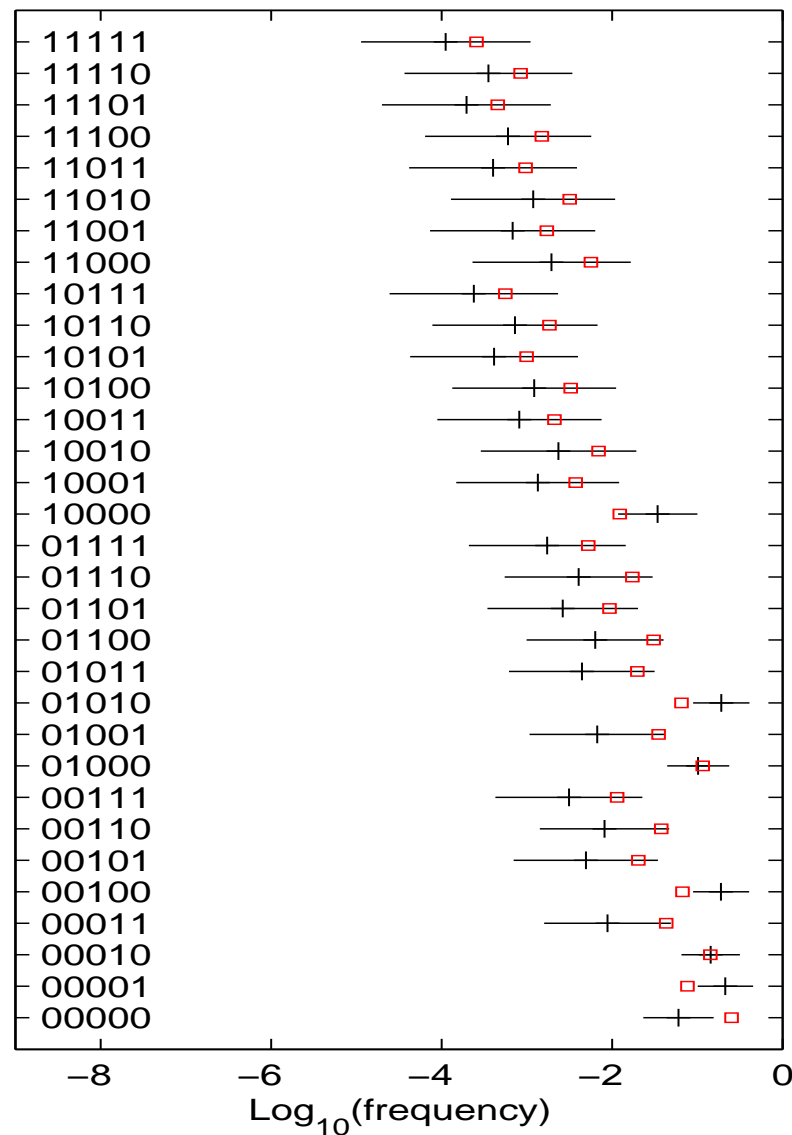


Estimated probabilities in cases

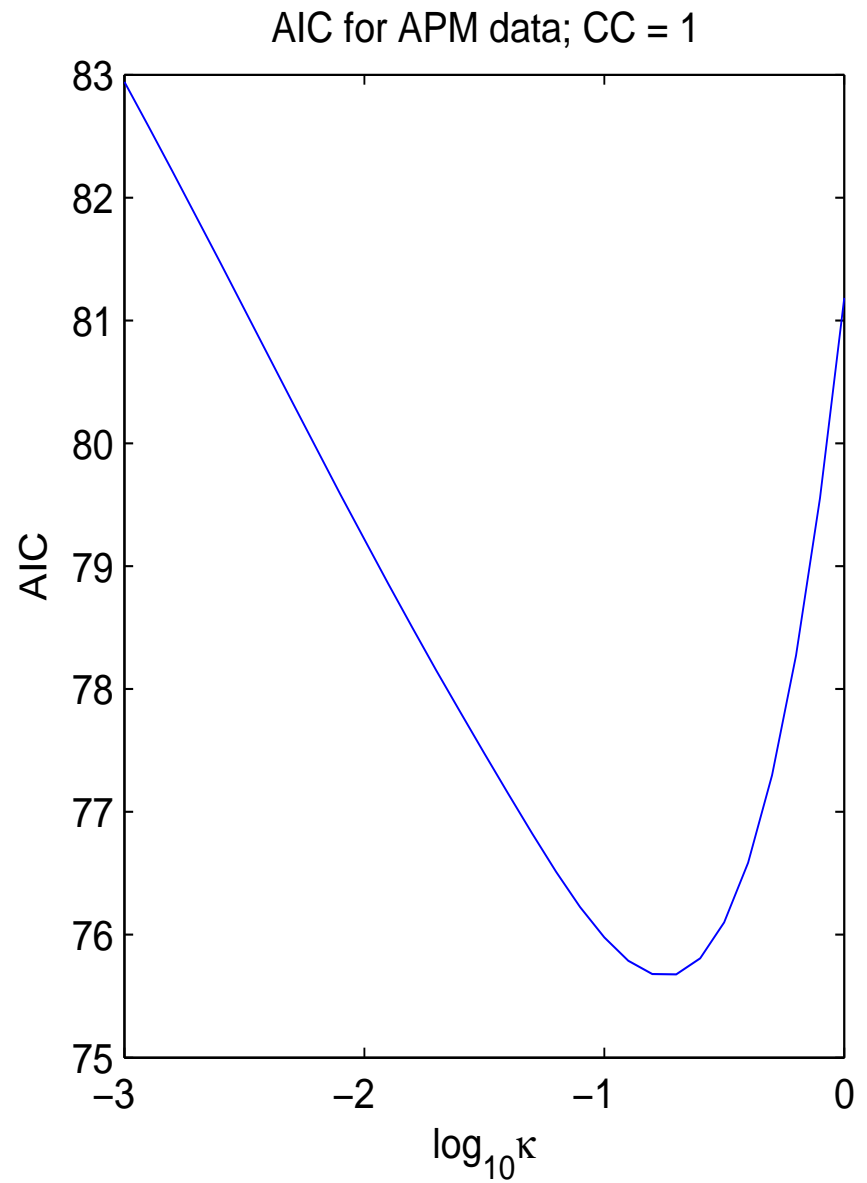
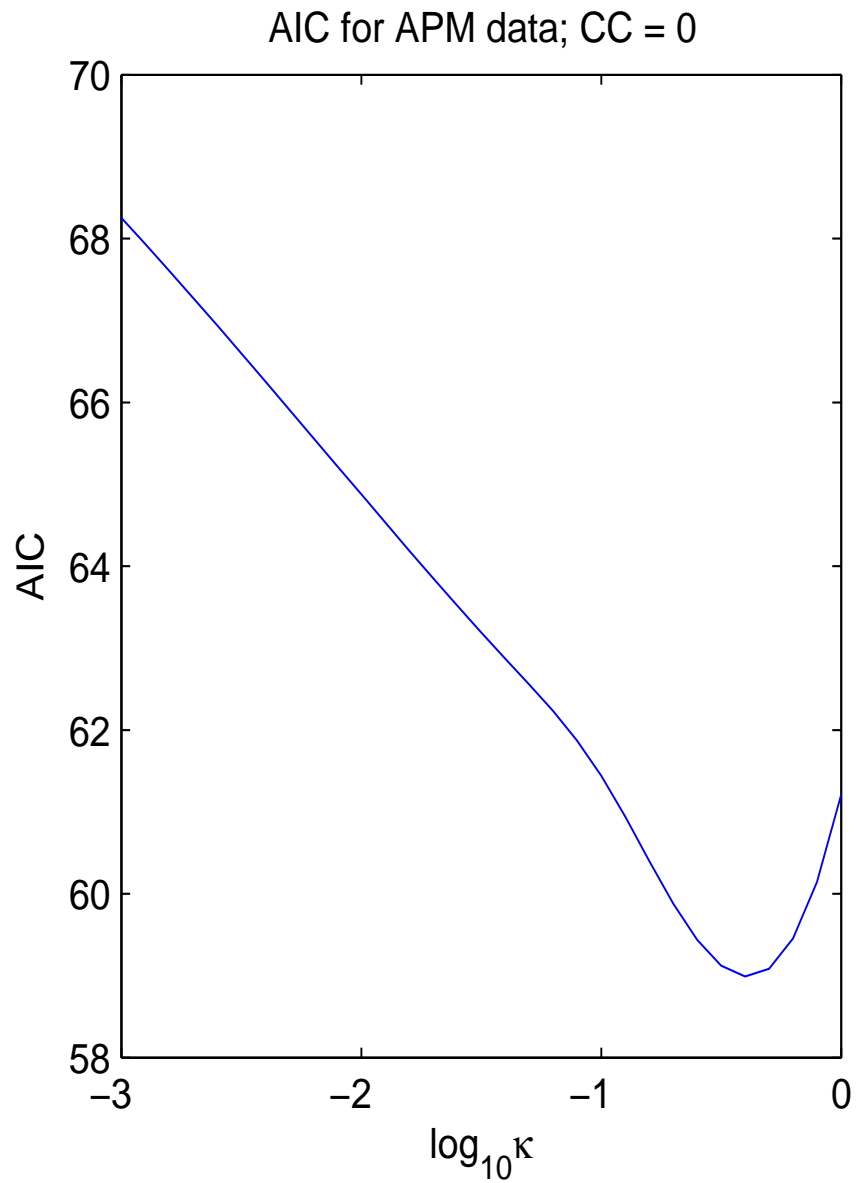
Estimates for $\kappa = 0.1$; CC = 1



Estimates for $\kappa = 1$; CC = 1



Selection of κ in cases and controls



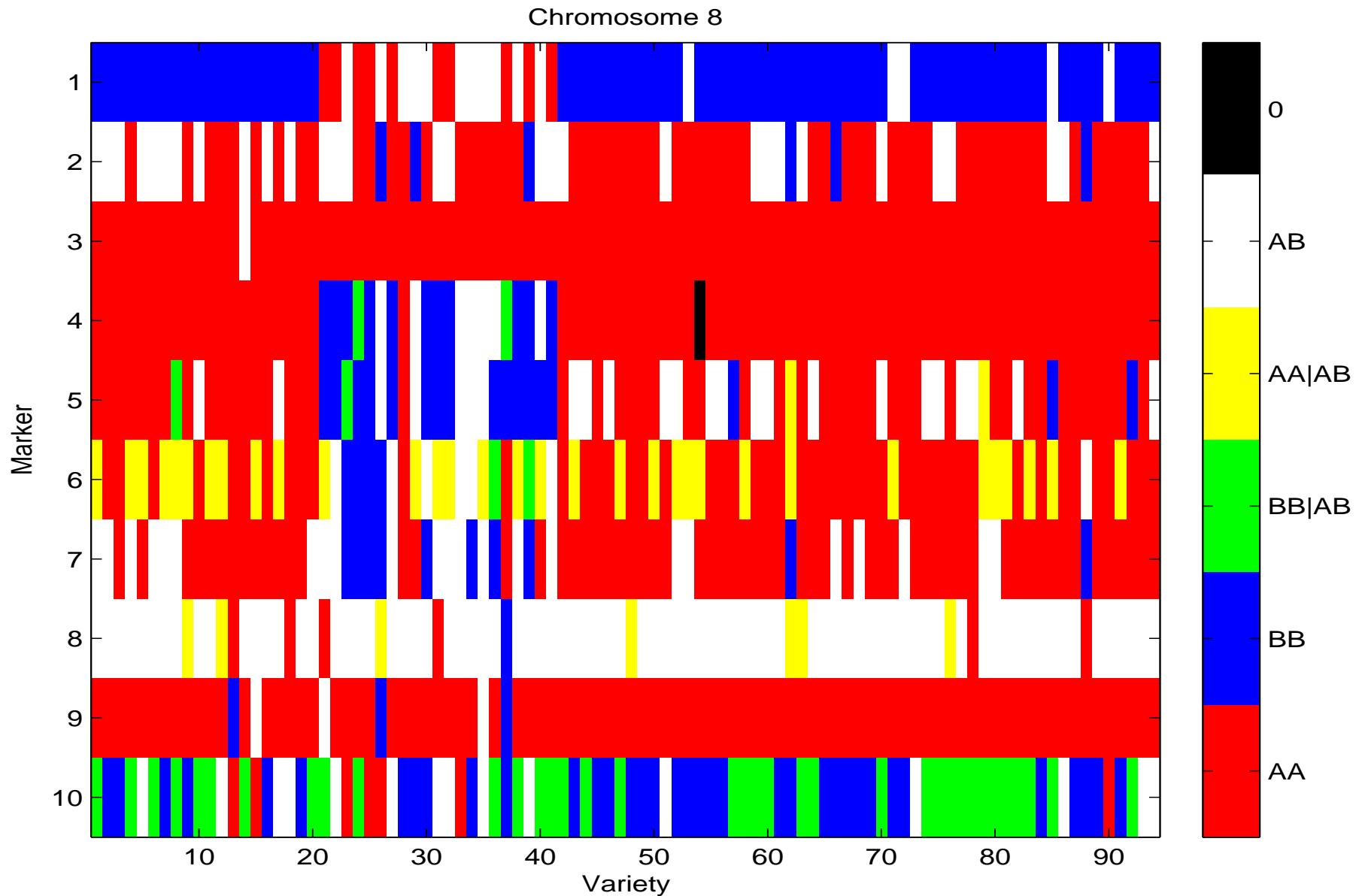
An alternative approach

- We modeled expected values of all genotype frequencies (in y)
- C has 3^L rows, but many genotypes are not observed
- Alternative: one row per individual
- Now μ gives probability of individual genotype
- Log-likelihood $L = \sum_i \log \mu_i - \lambda \sum_j \gamma_j$
- Second term for condition $\sum_j \gamma_j = 1$
- Lagrange multiplier λ (turns out to be m)
- When we have m observed individual genotypes

Complexities with AFLP markers

- Modern SNP technology is quite reliable
- We usually can trust genotypes
- In plant genetics (in Wageningen) AFLP is popular
- AFLP: amplified fragment length polymorphism
- Imperfect scoring; genotypes cannot be determined precisely
- Results may be $AA = 0, AB = 1, BB = 2$ (“crisp”)
- Or $AA/AB = 0/1$ or $AB/BB = 1/2$, or unknown = $0/1/2$

Example of fuzzy data: tomatoes



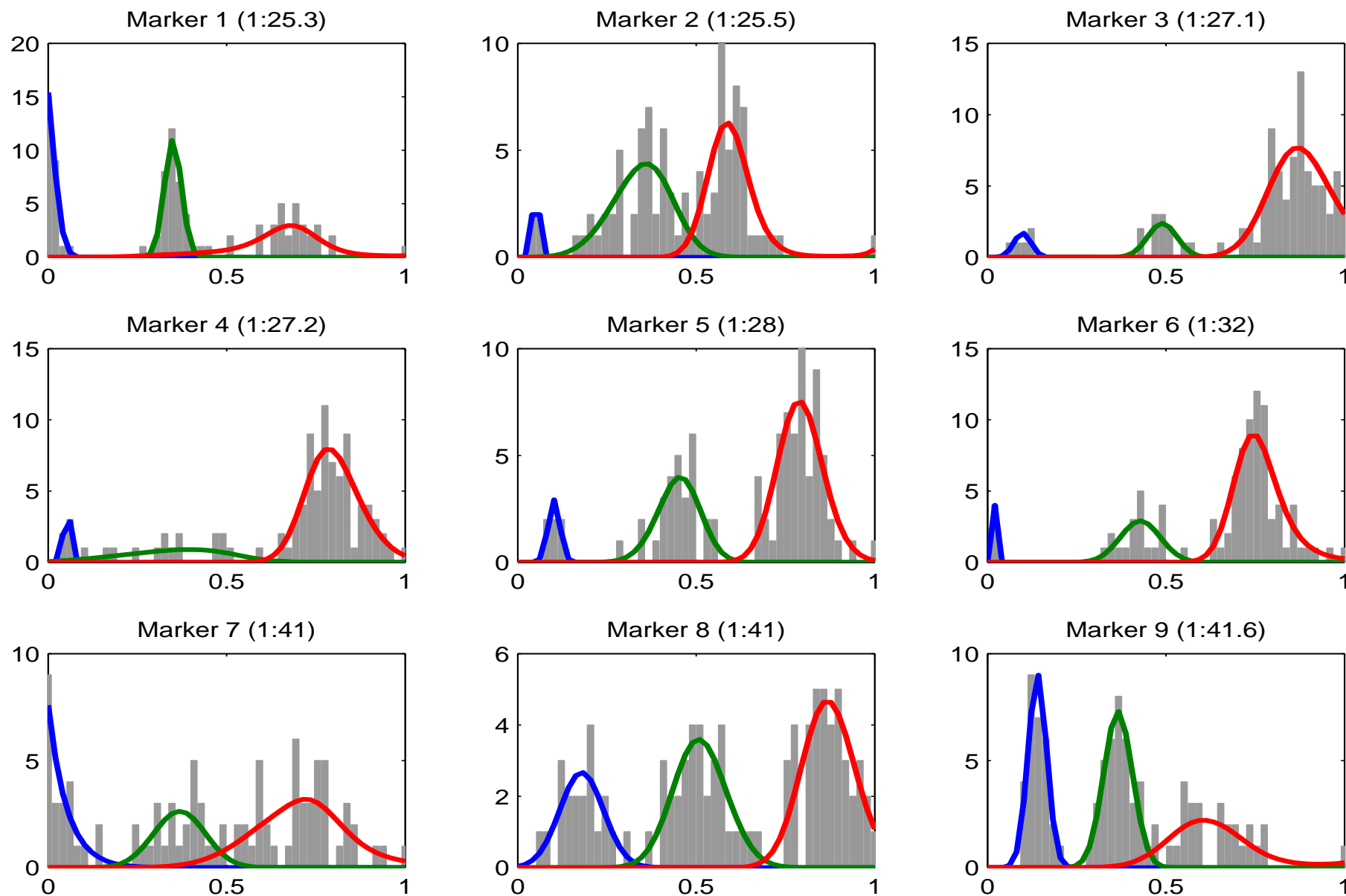
Coding vectors for one marker

- Vectors for crisp genotypes of one marker:
 $AA = [1\ 0\ 0]$, $AB = [0\ 1\ 0]$, $BB = [0\ 0\ 1]$
- Vectors for fuzzy genotypes:
 $AA/AB = [1\ 1\ 0]$, $AB/BB = [0\ 1\ 1]$, unknown = $[1\ 1\ 1]$
- Connect to each marker and individual such a vector
- Multiple markers: Kronecker products
- Collect them in rows of the *fuzziness* matrix F
- F has m rows and 3^L columns
- New composition matrix $C^* = FC$

AFLP mixtures

- The fuzziness vectors were relatively crisp
- Each genotype was either possible or not
- In real life we may need probabilities
- Derived from observed mixtures
- This leads to “soft” fuzziness vectors
- Or to a transition matrix of classification probabilities

Non-parametric AFLP mixtures in tomatoes



Things to work on

- It is hard to visualize fit with individual data
- Practical limit: 10 to 12 markers
- One solution: moving window for more than 10 markers
- We can use the model for (penalized) LD estimation
- Take all marker pairs, fit, compute LD measure
- My suggestion: deviance $d = \sum_k e^{\hat{\beta}_k} (\hat{\beta}_k - \alpha_k)$
- Can we combine haplotype and mixture estimation?
- Tetraploids and beyond?