Sequential Importance Sampler To Estimate Relatedness Structures

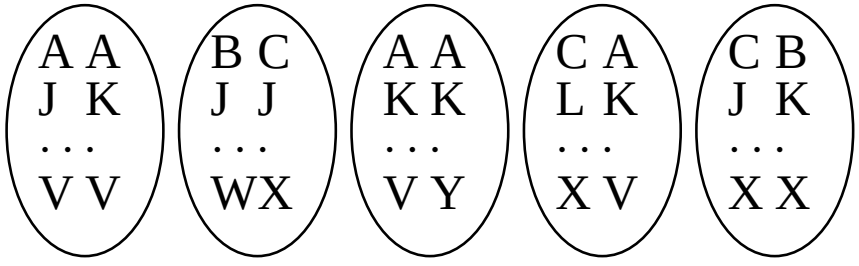
Toby Johnson

toby@maths.leeds.ac.uk

Centre for Statistical Bioinformatics, University of Leeds

(joint work with Kevin Dawson)

Estimating relatedness structure

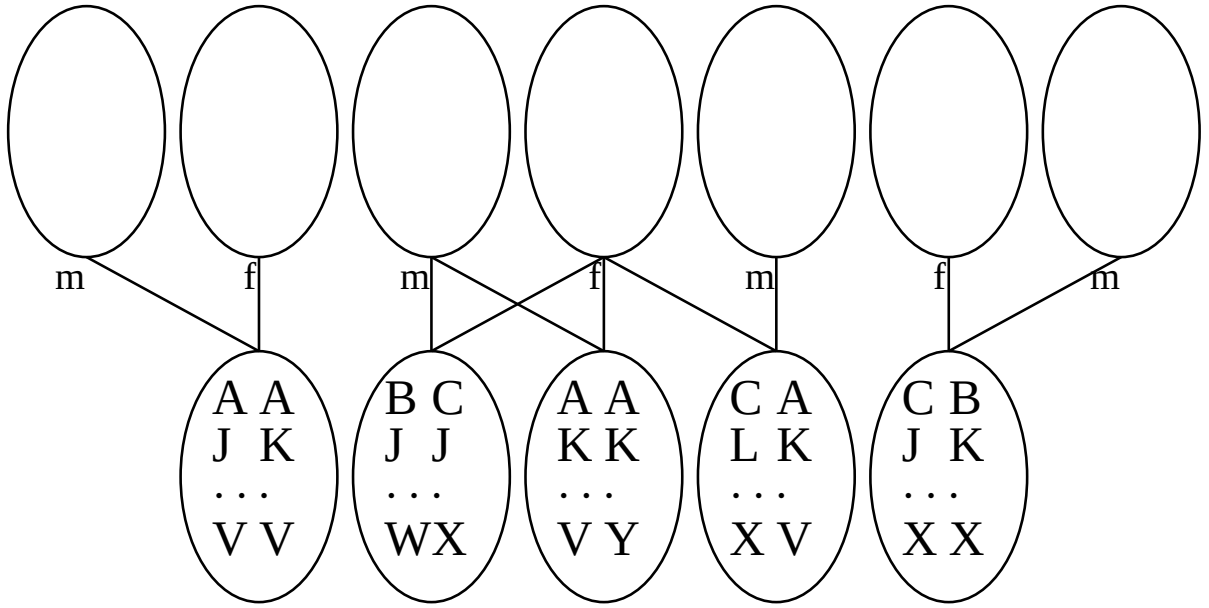


- Sample of n **individuals** (animals, plants)
- Each individual genotyped at ℓ **variable loci** (microsatellites; STRs)

  Each individual carries two alleles at each locus

  Alleles at locus 1 are A,B,. . . at locus 2 are J,K,. . . , . . . , at locus ℓ are V,W,. . .

Estimating relatedness structure



- Sample of n **individuals** (animals, plants)
- Each individual genotyped at ℓ **variable loci** (microsatellites; STRs)

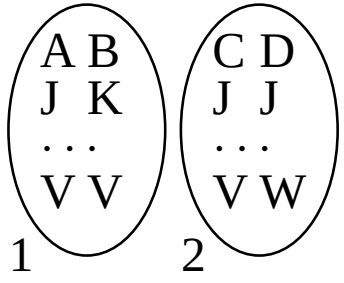
  Each individual carries two alleles at each locus

  Alleles at locus 1 are A,B,. . . at locus 2 are J,K,. . . , . . . , at locus ℓ are V,W,. . .
- We wish to **infer the pedigree**
  - May extend back in time several generations
  - Sampled individuals may not be contemporaneous

Motivation for estimating relatedness

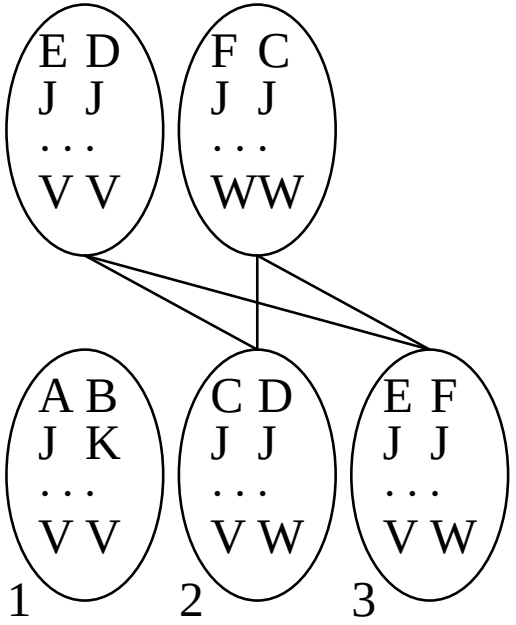
- For many questions of medical or agricultural importance, relatedness is directly observed (partially or completely)
- Inferring relatedness more important in study of **wild populations**, pests, parasites, disease vectors etc.
- Inference of relatedness allows inference of any quantity that could be inferred, if the relatedness structure had been directly observed. E.g.
  - Variance in any measured trait can be decomposed into genetic and non-genetic components; **heritability**
  - Parent–offspring **dispersal**
  - Distribution of parental **reproductive success**
  - Unbiased estimation of population allele frequencies etc. (relatives in sample cause duplicate counts, applications to case/control)
- Equivalent to making exact inference backwards one time step for Wright–Fisher/Cannings model
  - Analogue of making inference under the ARG, but discrete time, large sample $n \ll N$, free recombination

Motivation for whole sample based inference



- Pairwise estimation methods are very popular
- Suppose we want to assess the probability that individuals 1 and 2 are siblings

Motivation for whole sample based inference



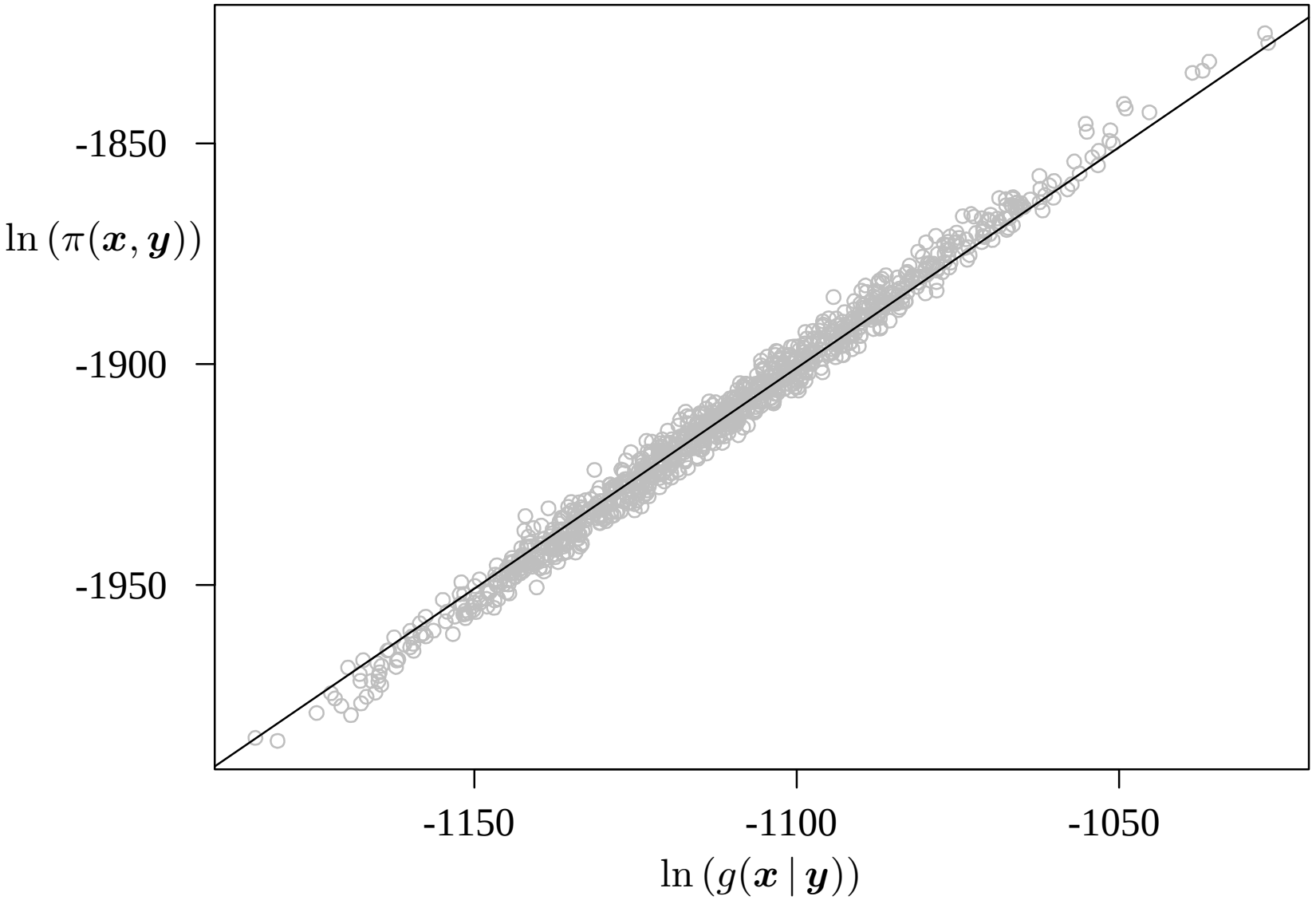
- Pairwise estimation methods are very popular
- Suppose we want to assess the probability that individuals 1 and 2 are siblings
- The presence of individual 3, who is a sibling of 2 with high probability, implies that 1 cannot be a sibling of 2 or 3 (**exclusion principle**)

General framework

- Candidate **parents arbitrarily labelled** $1, 2, \ldots, N$ (with N uncertain)
- Make **joint inference about allocation labels**, $\boldsymbol{a}$
  where the pair $(a_i^{\mathrm{m}}, a_i^{\mathrm{f}})$ indicates mother and father of i-th individual
- Some candidate parents may be genotyped
- Genotype frequencies in candidate parents are uncertain
- Relevant observational data may be available
- Distribution of pedigrees, $\pi(\boldsymbol{a} \mid N)$, may depend on unknown parameters

- Most/all candidate parents genotyped
  - Vanilla MCMC schemes work well (e.g. Hadfield *et al.* 2006)
- Few/no candidate parents genotyped, some methods proposed
  - MCMCMC (`parentage`; Emery, **Wilson** *et al.* 2001)
  - MCMC (`sibship`; **Dawson & Belkhir**)
  - Maximum likelihood approaches (Thomas & Hill 2000, 2002, Wang 2004)

Monte Carlo vs. maximum likelihood



Monte Carlo vs. maximum likelihood

- Maximum likelihood approach
  - Inference based only on most likely pedigree is only reliable with large family sizes (Thomas & Hill et al.)
  - Pedigree is not object of ultimate interest, “two step procedure is never right” (Hill)
  - Fast procedure available under better models
- Previous MCMC approaches
  - Slow and poor mixing for large datasets
  - Wilson MCMCMC seems like “sledgehammer”
  - Dawson–Belkhir approach relies on specially designed Metropolis–Hastings proposal; how general is this?
- Appeal of importance sampling approach
  - “Data savvy”
  - Easier to adapt to different models and priors
  - Less tuning may be required
  - Convenient estimate of Monte Carlo error
- Plan is to focus on simplest model, and get sampling working well

(Almost) the simplest model

- Number of candidate parents, N, known
- Parental genotype probabilities known and independent across loci and parents

Genotype	Probability
AA	$p_{\mathrm{A}}^2 + f_j p_{\mathrm{A}}(1-p_{\mathrm{A}})$
AB	$(1-f_j)p_{\mathrm{A}}p_{\mathrm{B}}$
BA	$(1-f_j)p_{\mathrm{B}}p_{\mathrm{A}}$
BB	$p_{\mathrm{B}}^2 + f_j p_{\mathrm{B}}(1-p_{\mathrm{B}})$
...	...

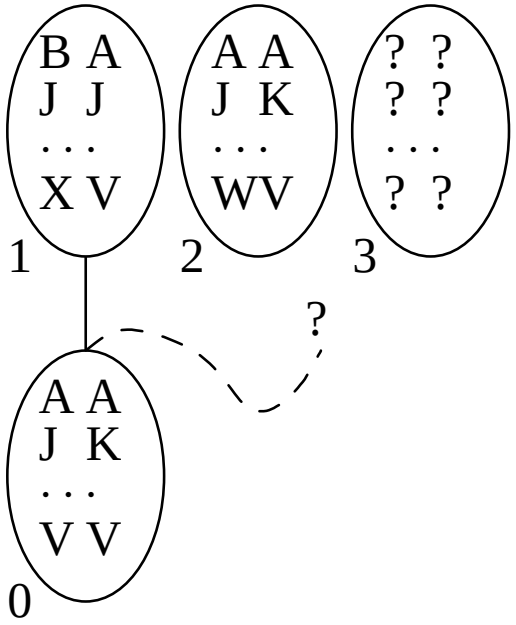
  where p_{A} are allele frequencies and f_j is identity-by-descent at j-th locus

- Simple prior on pedigree; parental assignments i.i.d. with

$$\pi(a_i^{\mathrm{m}}, a_i^{\mathrm{f}}) = \begin{cases} s\frac{1}{N} & \text{if } a_i^{\mathrm{m}} = a_i^{\mathrm{f}} \\ (1-s)\frac{1}{N}\frac{1}{N-1} & \text{otherwise} \end{cases}$$

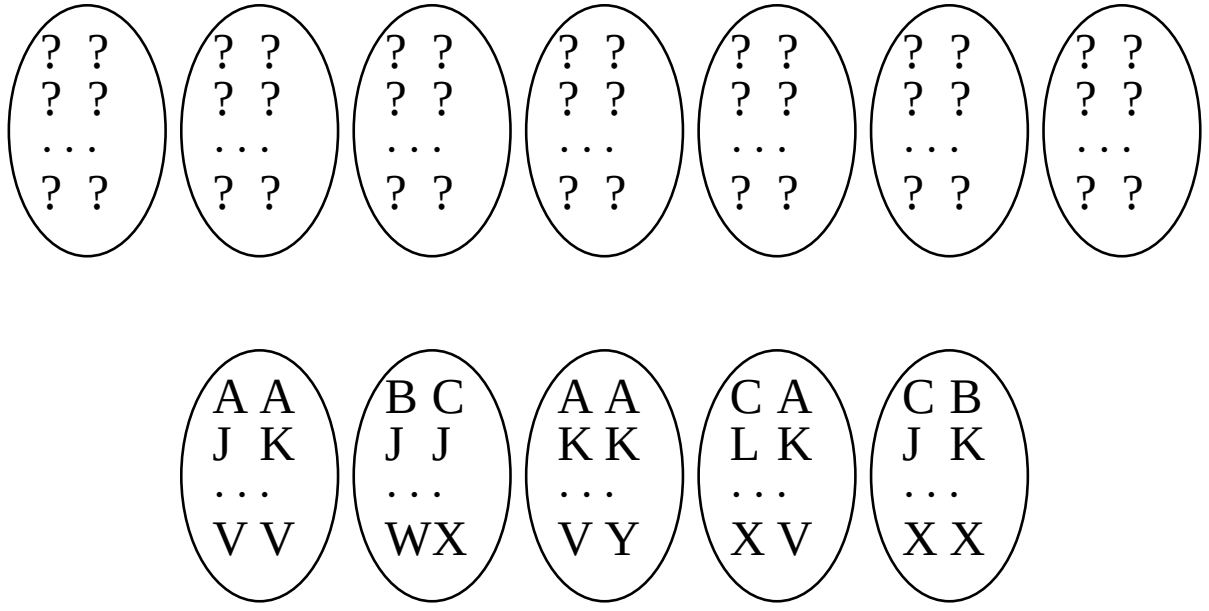
  with self-fertilisation rate s assumed known

What can we already do?

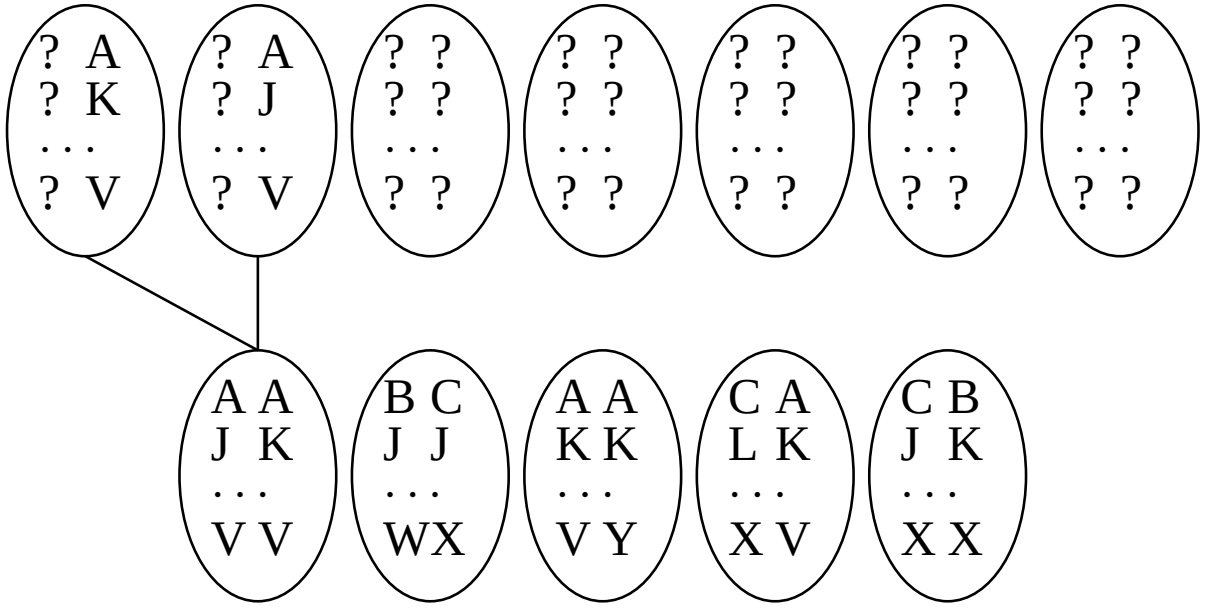


- $\Pr\left(y_0 | a_0^{\mathrm{m}} = 1, a_0^{\mathrm{f}} = 2, y_1, y_2\right) = 1 \times \frac{1}{2} \times \cdots \times \frac{1}{2}$
- $\Pr\left(y_0 | a_0^{\mathrm{m}} = 1, a_0^{\mathrm{f}} = 3, y_1, y_3\right) = p_{\mathrm{A}} \times p_{\mathrm{K}} \times \cdots \times p_{\mathrm{V}}$
- These are the “match probabilities” used in forensics
- Can be computed for any amount of missing data
- With a prior, we can sample (both) the parental assignments, given any genotype data

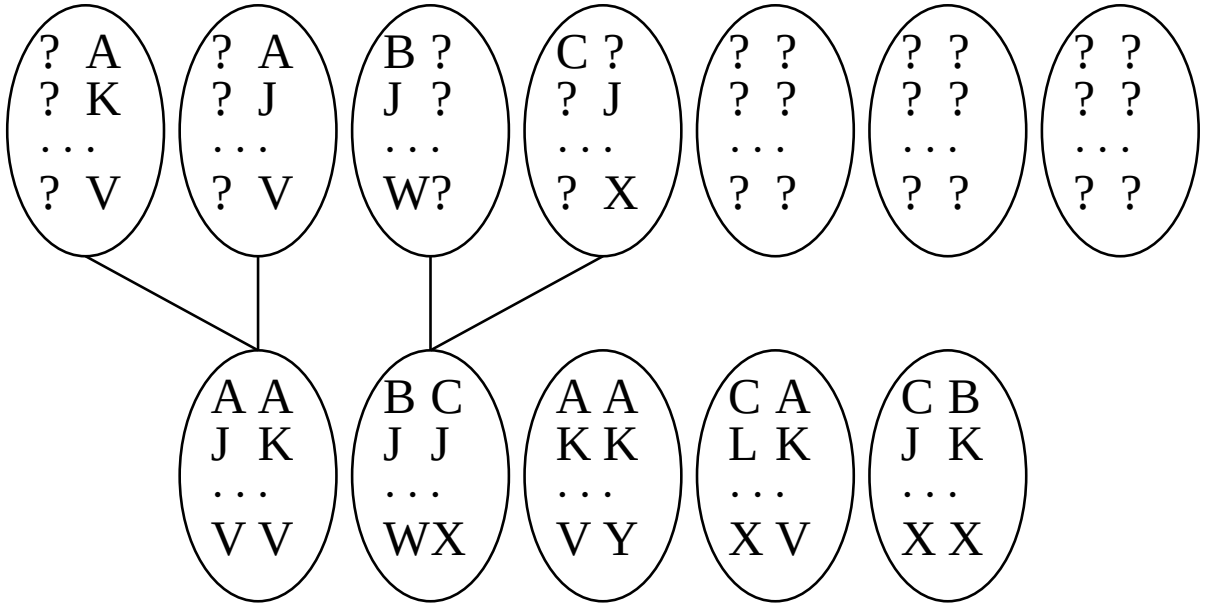
Sequential sampling/imputation



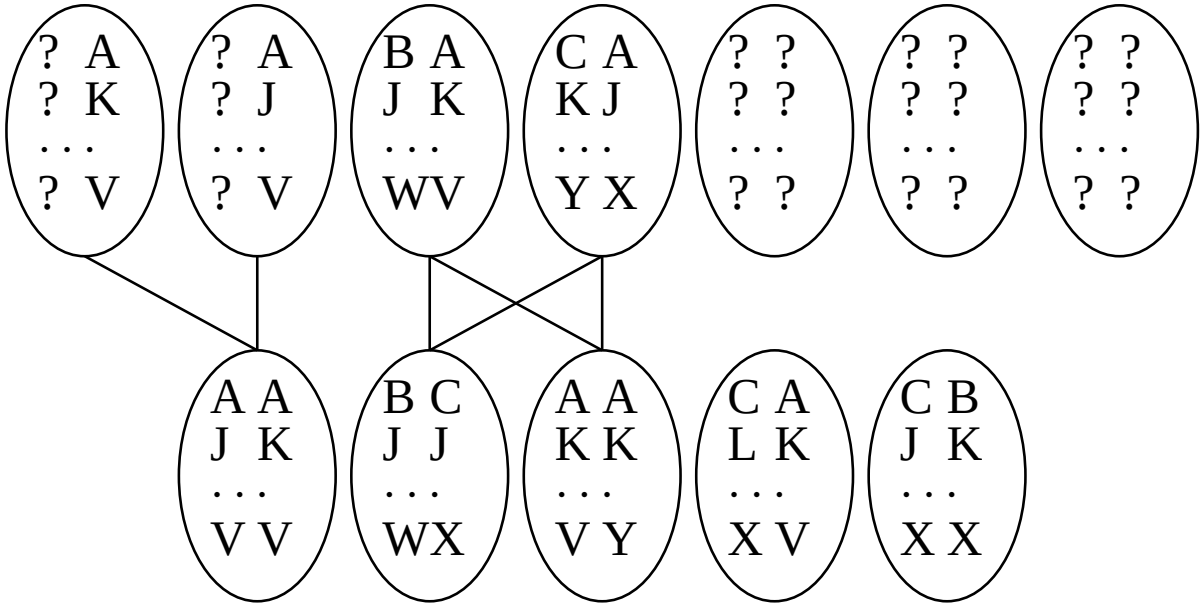
Sequential sampling/imputation



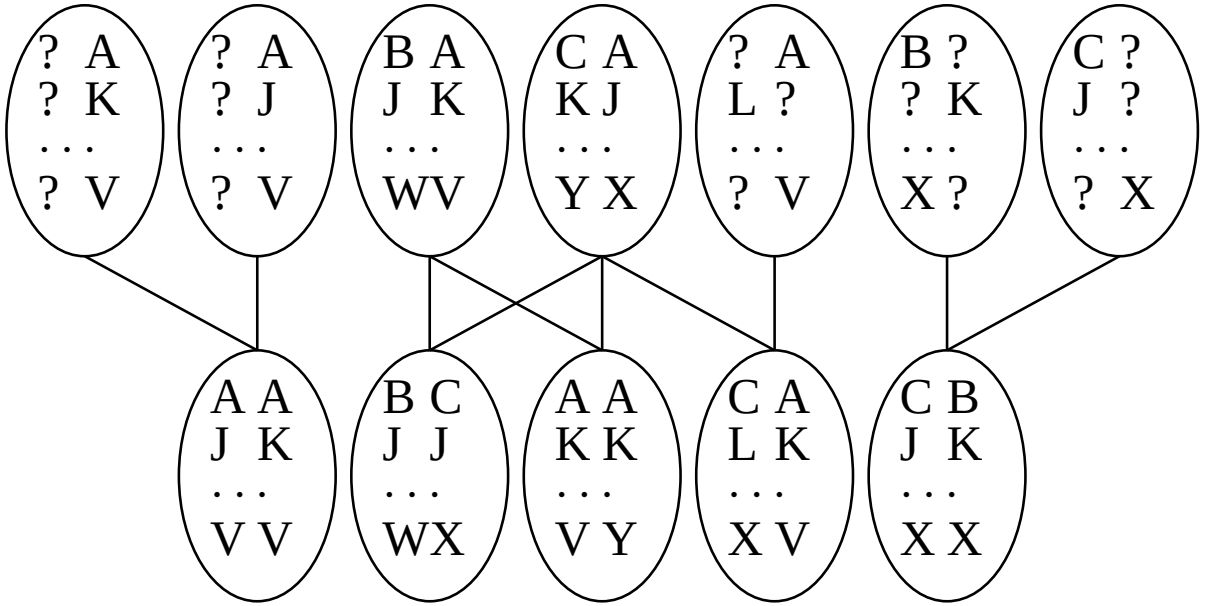
Sequential sampling/imputation



Sequential sampling/imputation

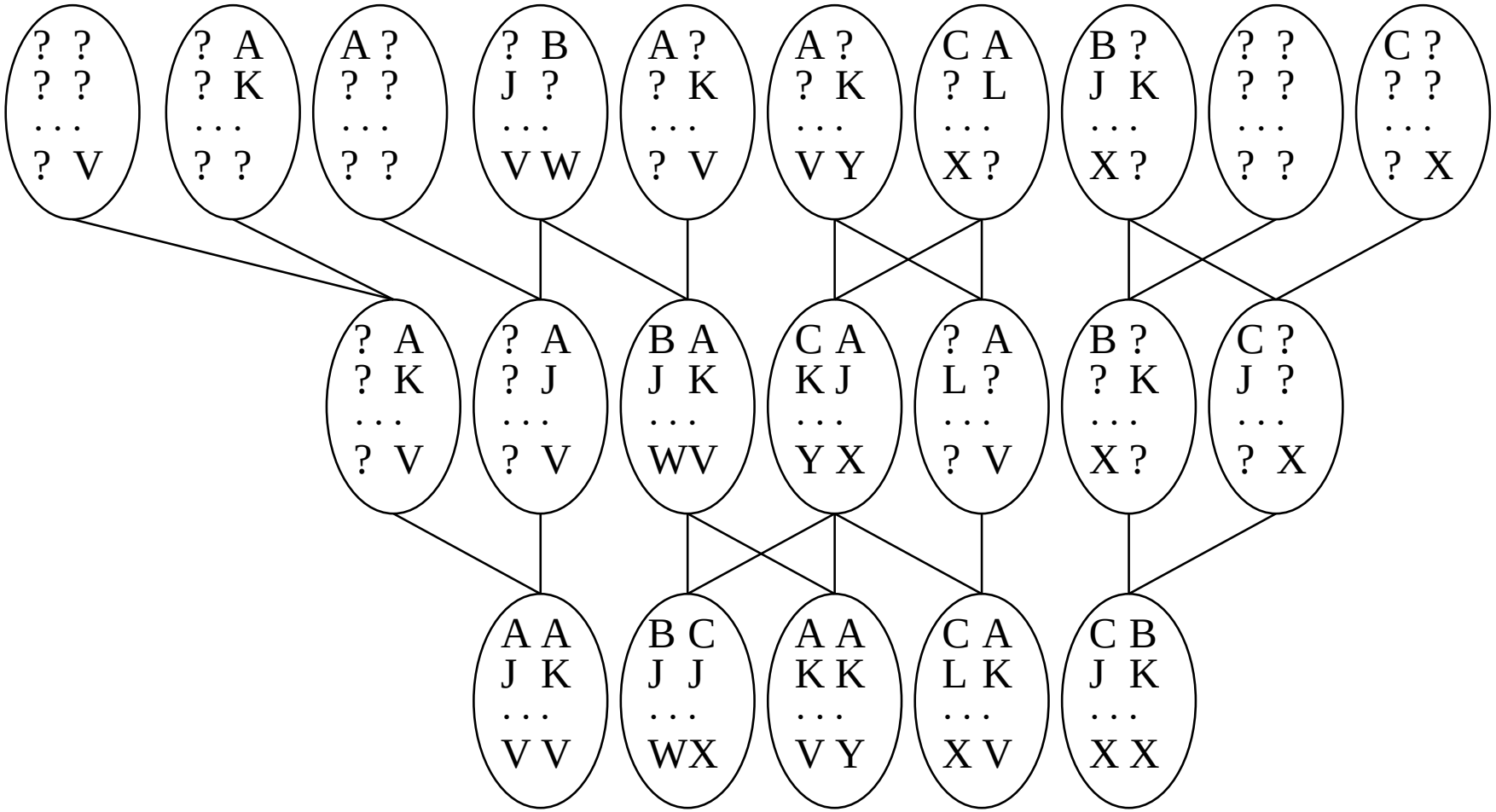


Sequential sampling/imputation



- We can **sequentially sample parents**, **randomly imputing** parental **genotypes** as we go
- We are **pretending** to observe the sampled individuals sequentially
  can still **target the full data non-sequential posterior**, but may do so inefficiently
- Simple to include known parentage and known candidate parent genotypes
- Connection with PAC model of Li & Stephens

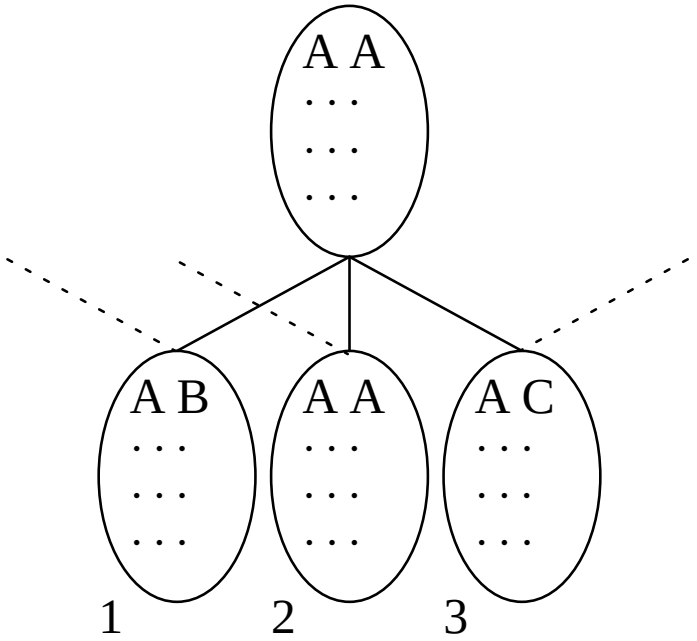
Deeper pedigrees



- Potential for second sequential aspect, going backwards in time…

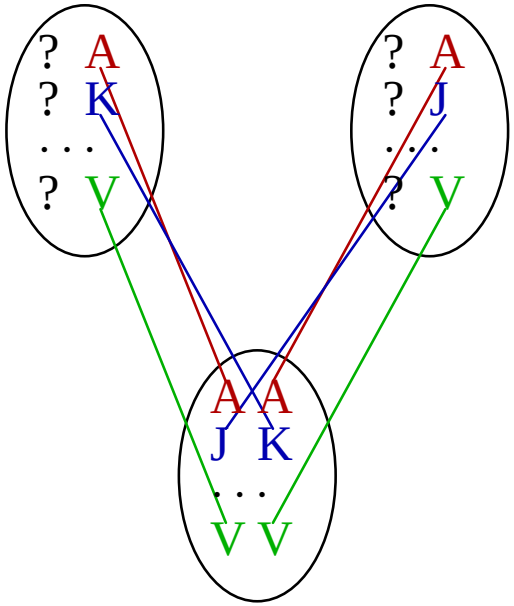
The first problem

- The “obvious” space is
  pedigree $\otimes$ order of sampled individuals $\otimes$ parental genotypes
- A serious problem is that the proposed sequential sampling/imputation scheme can hit the same sample **in several different ways**



- It will however turn out that is the “right” proposal mechanism

Descent graphs

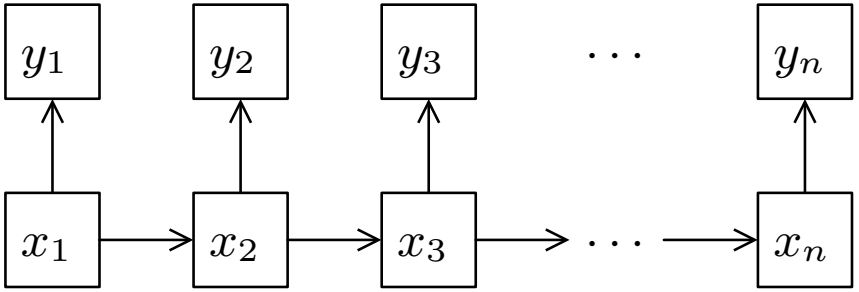


- Let g_i denote the descent graph for the i-th sampled individuals
- Let g_{ij} indicate the subgraph for the j-th locus
- For two distinct parents there are 8 *a priori* equally probable such subgraphs, independent across loci
- For one distinct (self-fertilising) parent there are 4 *a priori* equally probable such subgraphs, again independent across loci

Recasting problem as dynamic model

- The order o is a (uniform) random permutation of $(1, 2, \ldots, n)$
- Let y_i be the **observed genotype** of the o_i-th offspring
- The **assignment to be inferred** is $x_i = \{a_i^{\mathrm{m}}, a_i^{\mathrm{f}}, g_{i1}, \ldots, g_{i\ell}, y_i^*\}$

  where y_i^* is (perhaps partial) parental genotypes including all alleles pointed to by **all descent graphs $g_1, g_2, \ldots, g_i$ up to the i-th one**.
- Then we have a dynamic model:



  (this description makes the problem seem **boring** and **hard**)
- Remark: We will always use probabilities marginal to parental genotypes that are not transmitted to any sampled offspring

The naïve sequential importance sampler (SIS)

- We know the "guiding" **series of densities**

$$\pi_1(x_1, y_1), \pi_2(x_1, y_1, x_2, y_2), \ldots, \pi_n(x_1, y_1, \ldots, x_n, y_n)$$

- Final density is (proportional to) the target posterior
- These give us a series of **proposal densities**

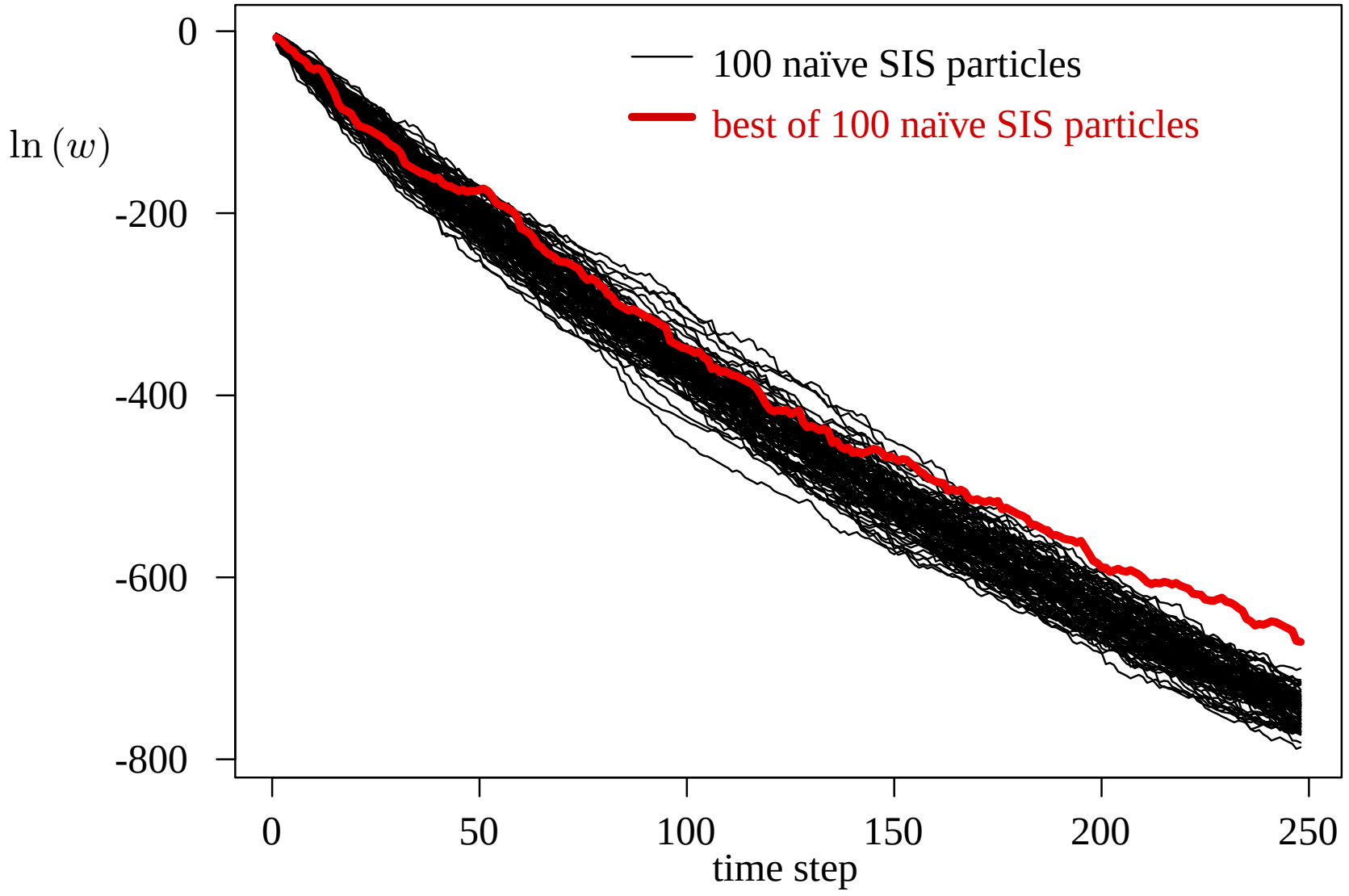
$$g_1(x_1 \mid y_1), g_2(x_2 \mid x_1, y_1, y_2), \ldots, g_n(x_n \mid x_1, \ldots, x_{n-1}, y_1, \ldots, y_n)$$

  that we can sample from, and compute normalising constants for
- Importance weight of a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ is

$$w = \frac{\pi_n(x_1, y_1, \ldots, x_n, y_n)}{g_1(x_1 \mid \cdot) g_2(x_2 \mid \cdot) \cdots g_n(x_n \mid \cdot)}$$

- Note: $\mathrm{E}_{g_n}(w) = \pi(y_1, \ldots, y_n)$
- Importance weights can be calculated iteratively at every step
- Above is a special case where proposals are marginals of target series of densities but we can use construct the g using π for an approximating model

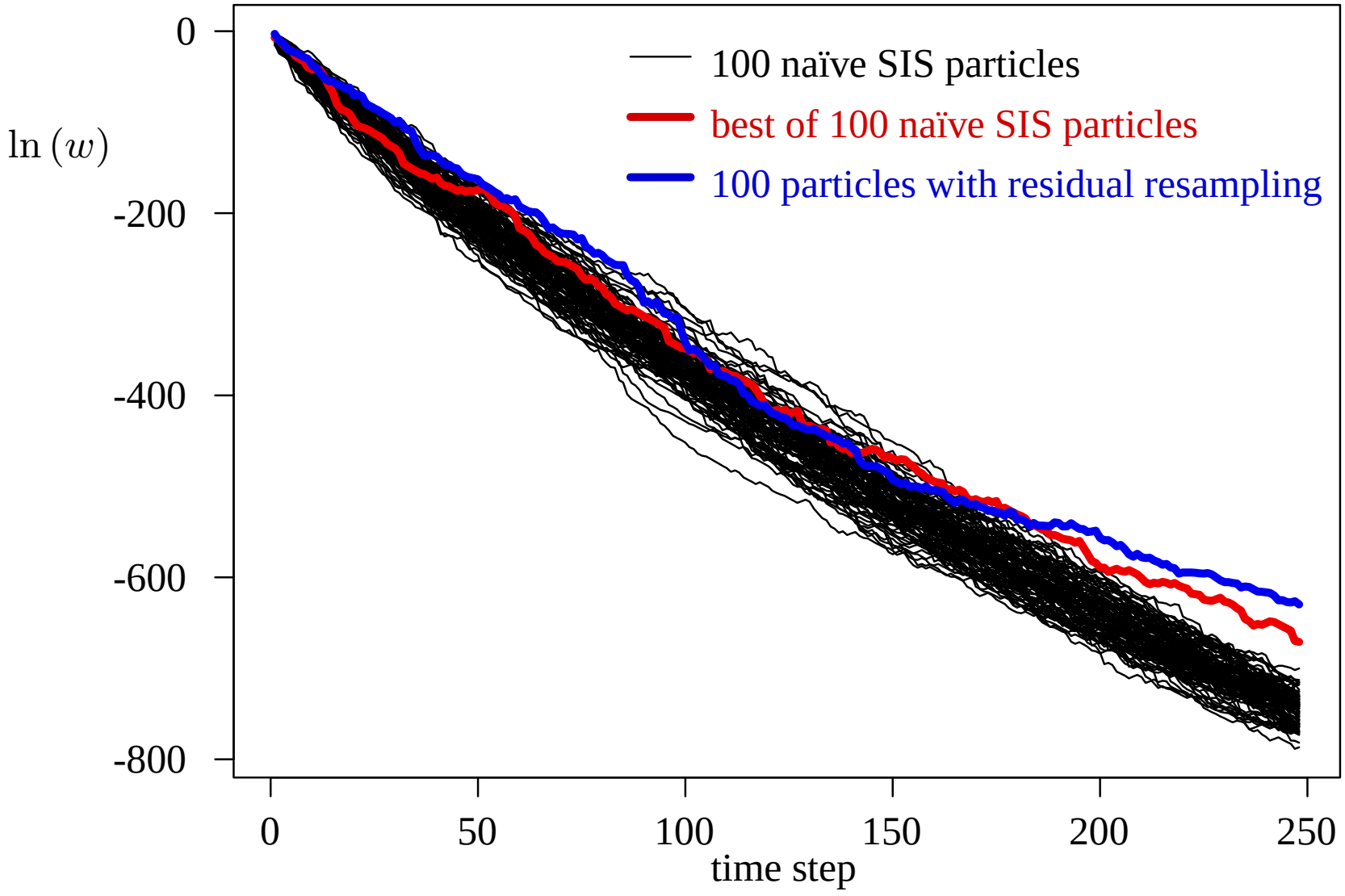
Performance of naïve SIS



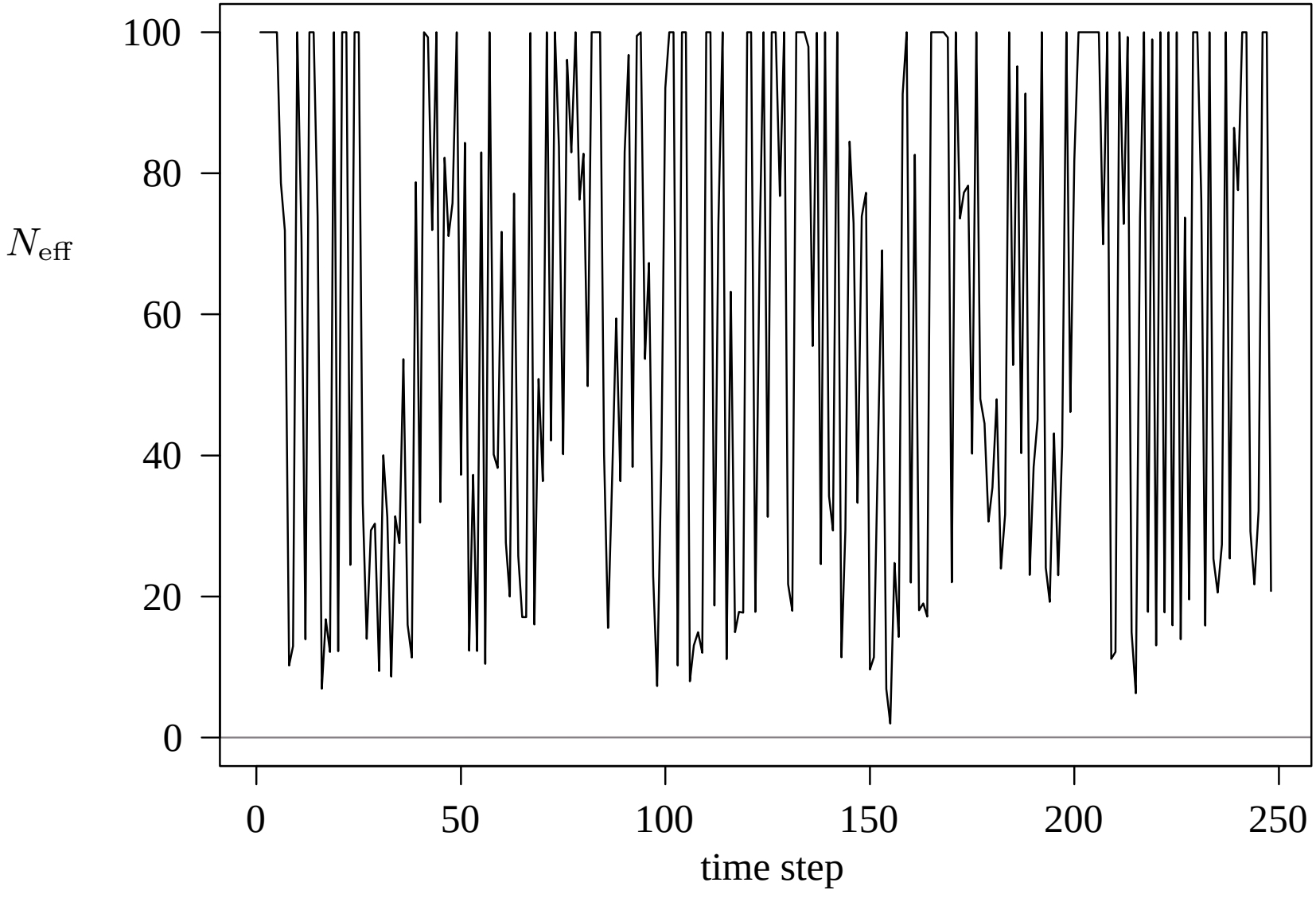
Residual resampling / particle filtering

- Residual resampling (see e.g. Liu 2001 p.72)
  - Suppose M particles with weights $w_1, w_2, \ldots, w_M$ (dependence on time step suppressed)
  - Compute normalised weights $\tilde{w}_i = w_i \big/ \left(\sum_i w_i\right)$
  - Retain $\lfloor M\tilde{w}_i \rfloor$ copies of i-th particle
  - Sample particles with replacement with probabilities proportional to "residual" $M\tilde{w}_i - \lfloor M\tilde{w}_i \rfloor$ until we have M particles
  - Reset all weights to $\left(\sum_i w_i\right) / M$
- This seems to work well only if all particles use the same order o of the observed data
  - Remember importance weights are running estimates of marginal likelihood of all data considered so far, $\mathrm{E}_{g_t}(w_t) = \Pr(y_1, \ldots, y_t)$
  - Want a scheme that adaptively finds sample orders that approximate the target density better; immune to this effect
- Remark: Multiple checkpoint rejection control an appealing alternative, because early proposals are cheaper to sample from

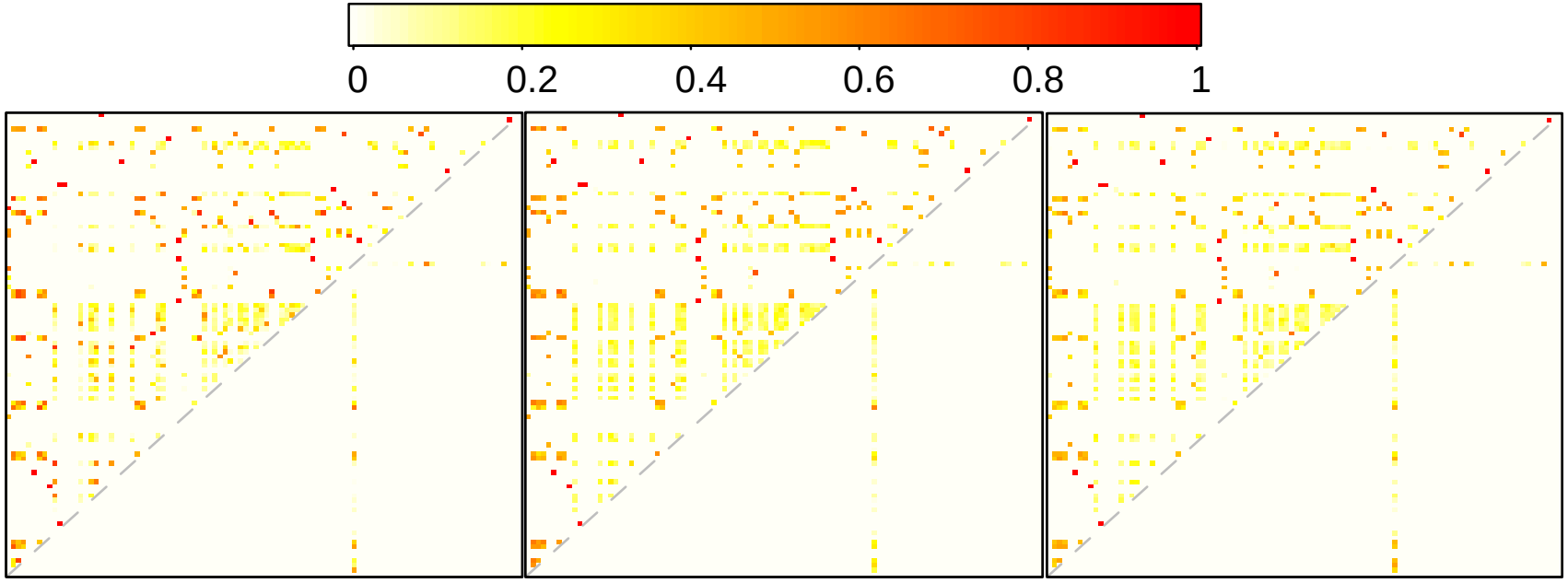
Example of residual resampling (SIR)



Effective number of particles with SIR



Effect of number of particles (for $n = 100$ slugs, $\ell = 6$)



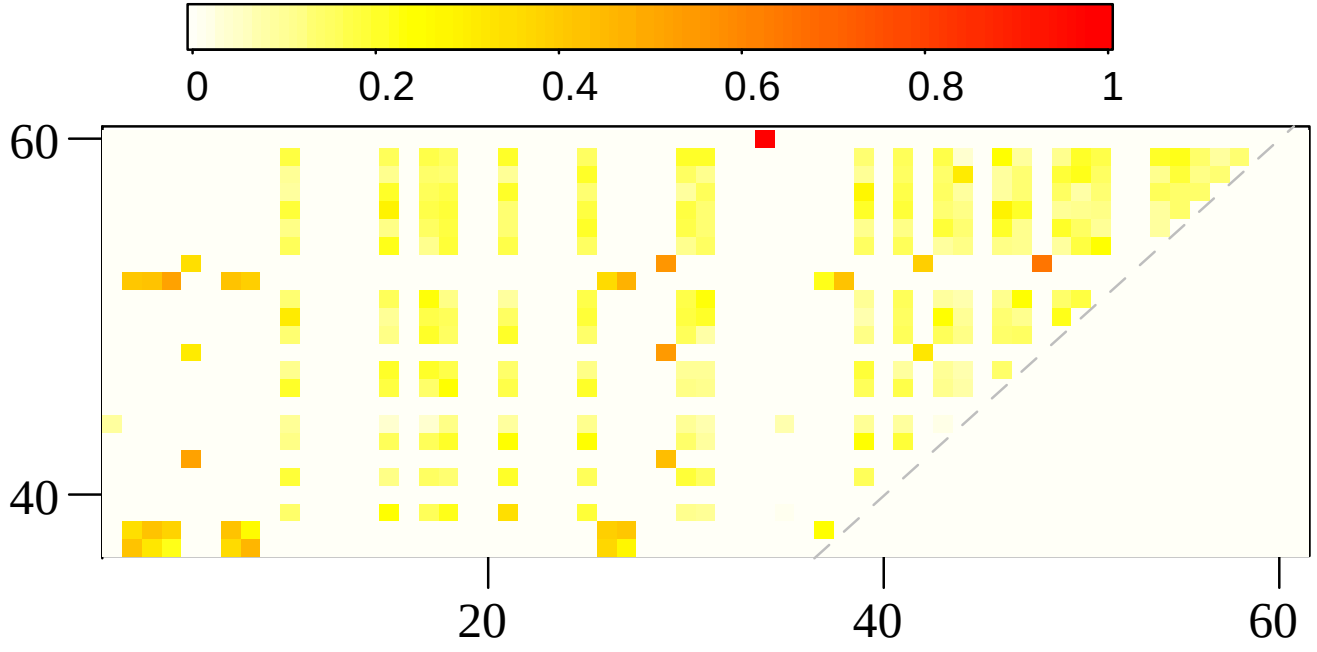
100 particles
$\ln(\mathrm{E}(w)) = -222.7$
$\mathrm{cv}(w) = 175.9$

500 particles
$\ln(\mathrm{E}(w)) = -219.9$
$\mathrm{cv}(w) = 25.0$

1000 particles
$\ln(\mathrm{E}(w)) = -218.0$
$\mathrm{cv}(w) = 46.4$

- Estimates using 500 replicate runs of each particle filter
- For 5000 particles $\ln(\mathrm{E}_g(w)) \simeq -217.6$
  so each particle has $\times 1.5$ greater weight at $\times 5$ cost

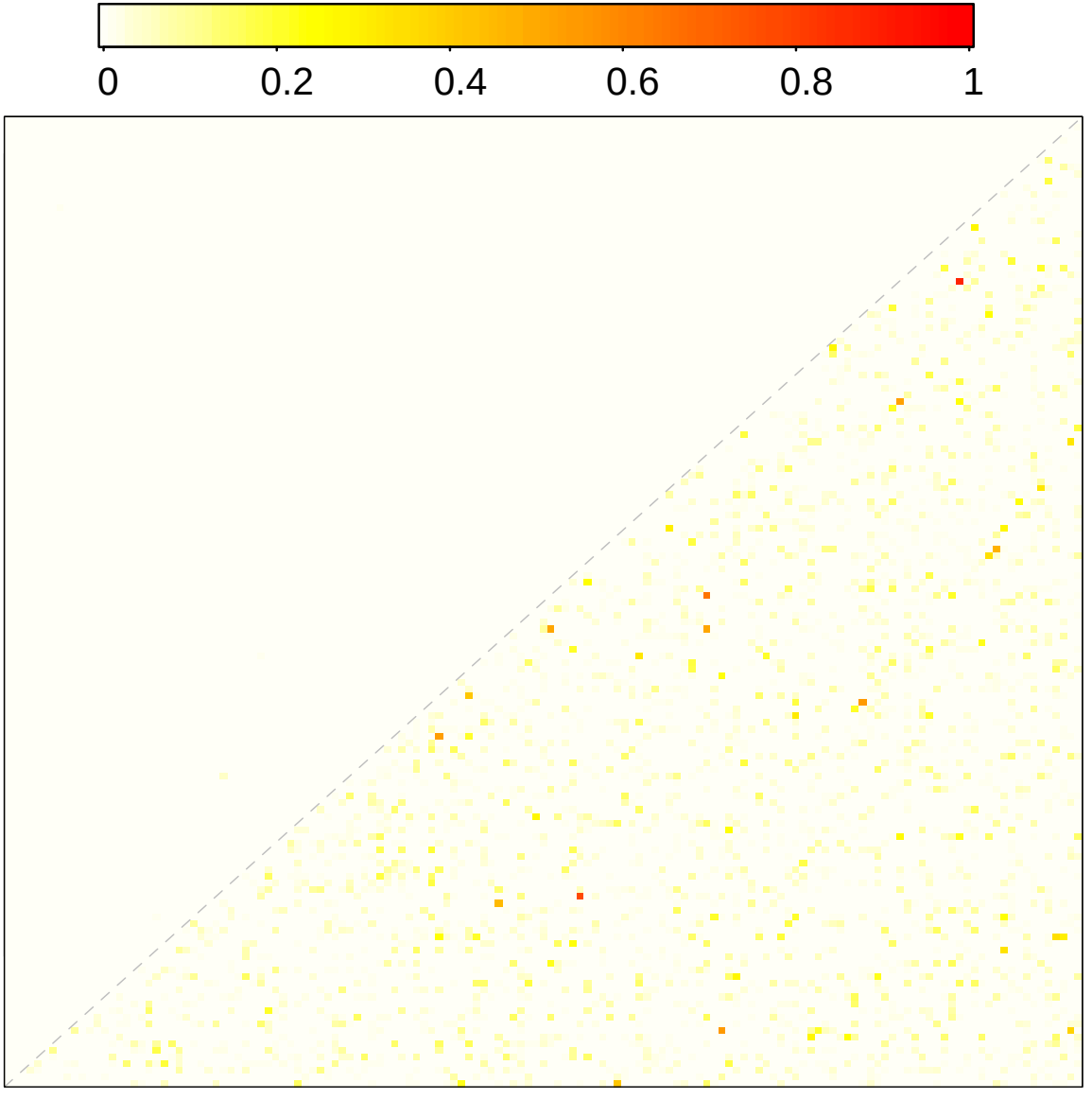
Some detail on posterior sibships



- Individuals with identical genotypes (should) produce symmetries in true posterior

Another application

(for $n = 145$ red deer, $\ell = 9$)



Concluding remarks

- Sequential Monte Carlo method showing potential to improve over MCMC methods
  - Better approach than independent runs of particle filter ?
  - SIMCMC ?
- Inference of relatedness sensitive to prior on pedigree / relatedness structure
- Need to use current machinery as importance sampler for posterior under better models, e.g.
  - guiding densities $\pi_t(x_{1:t}, y_{1:t})$ can be written down marginal to (unknown) allele frequencies
  - parental genotypes in partially selfing population