Maximum likelihood pedigree reconstruction using integer programming

James Cussens, University of York

WCB-10
21 July 2010

Outline

Pedigrees

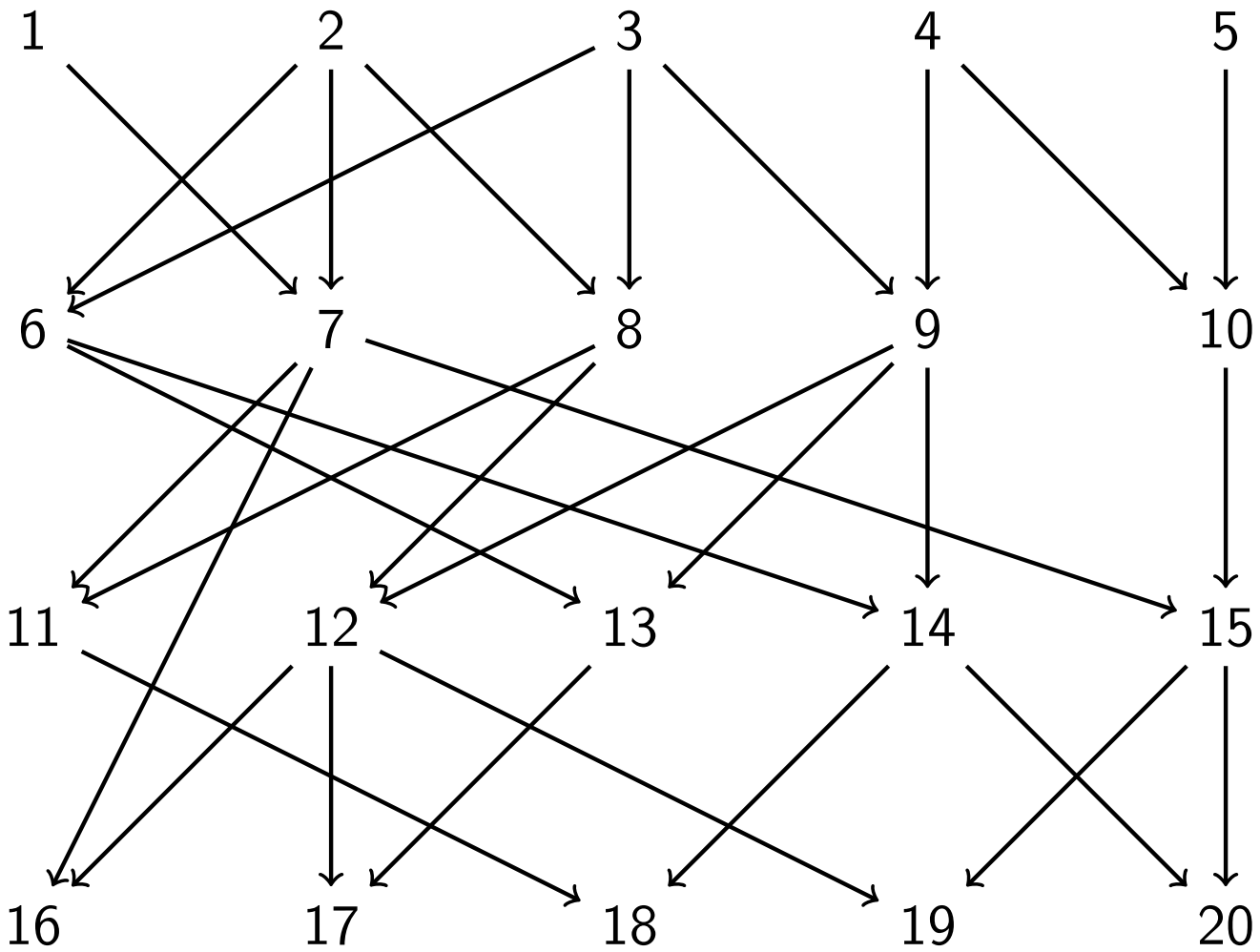
Integer linear programming

Encoding the pedigree problem

Results

Future directions

Pedigree of 20 individuals



Using genotypic data

- Individuals have (inherited) genetic differences.
- A location on a particular chromosome where the DNA varies between individuals is called a *marker locus*.
- For each individual assume we know which particular (pair of) DNA variants (*allele*) occurs at each marker locus: the individual's *genotype*.
- We assume complete genotypic data.

The likelihood function

- If j and k are the parents of i we can compute $P(G_i|G_j, G_k)$ using basic Mendelian genetics (where G_i is the genotype of individual i).
- Also assume we know allele frequencies in the general population then can get $P(G_i)$ for an individual with unknown parents.
- $P(\mathrm{data}|\mathrm{pedigree})$ is then a product of such probabilities, one for each individual.
- So $\log P(\mathrm{data}|\mathrm{pedigree})$ is a sum of 'scores' one for each individual.

Outline

Pedigrees

Integer linear programming

Encoding the pedigree problem

Results

Future directions

Integer linear programming

Maximise: $\mathbf{c}^T\mathbf{x}$
subject to $A\mathbf{x} \leq \mathbf{b}$

- $\mathbf{x} \in \mathbb{R}^n$, but some x_i constrained to be in $\mathbb{Z}$.
- $\mathbf{c}^T\mathbf{x}$ is the *objective function*
- $A\mathbf{x} \leq \mathbf{b}$ defines a *convex polytope*
- NP-hard.

Outline

Pedigrees

Integer linear programming

Encoding the pedigree problem

Results

Future directions

Encoding the graph with binary variables

Create $O(n^3)$ binary variables:

$$I(W \rightarrow v)(G) = \begin{cases} 1 & \text{if } v \text{ has parents } W \text{ in } G \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

The objective function

The ILP problem is then:

> Find an instantiation of the $I(W \rightarrow v)$ which maximises:
> $\sum_{v,W} \log P(G_v|G_W) I(W \rightarrow v)$
> subject to the $I(W \rightarrow v)$ representing a valid pedigree.

The $\log P(G_v|G_W)$ coefficients are pre-computed (using e.g. Robert Cowell's software).

Exactly one set of parents for each individual

$$\forall v : \sum_{W} I(W \rightarrow v) = 1$$

Ruling out cycles

- Create an integer-valued variable $\mathrm{gen}(v)$ for each $v \in V$.
- $\mathrm{gen}(v)$ is the length of the longest path from a *founder* (no known parents) to v in a pedigree.
- If j is a parent of i then $\mathrm{gen}(i) \geq \mathrm{gen}(j) + 1$.
- The following $O(n^2)$ constraints suffice to rule out cycles:

$$\forall u, v : \mathrm{gen}(v) - \mathrm{gen}(u) \geq 1 - n + n \sum_{W : u \in W} I(W \rightarrow v)$$

Ensuring sex-consistency

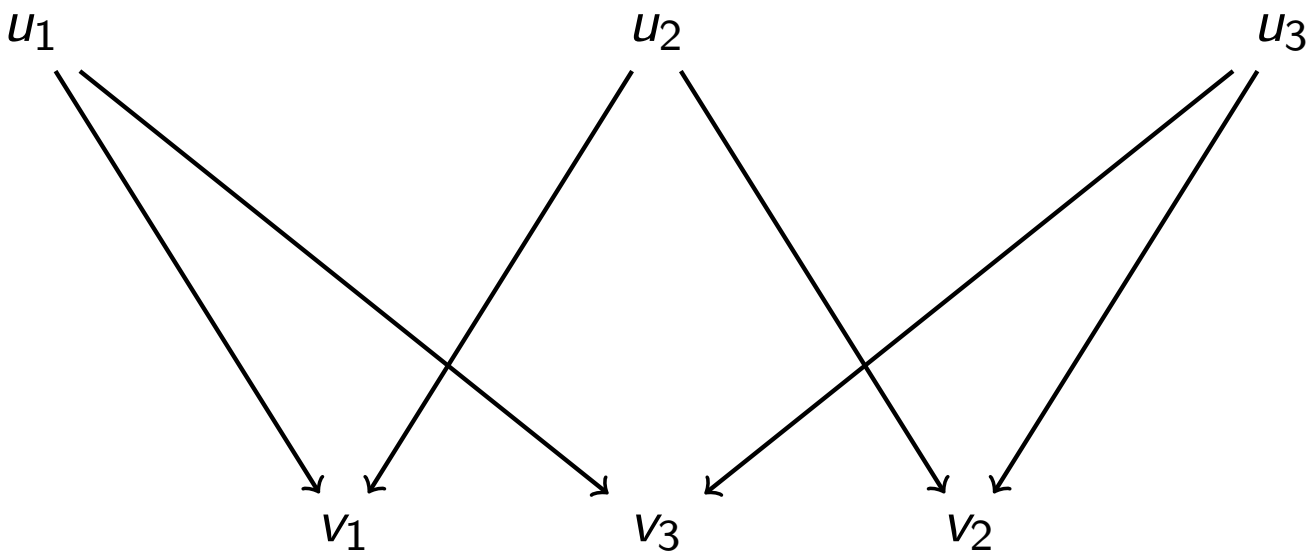


Figure: A sex-inconsistent pedigree. It is not possible to consistently assign a sex to each individual.

Ensuring sex-consistency

So let, $I_f(u)$ indicate that u is a female, then:
At most one mother:

$$\forall u, v, w : I(\{u, w\} \rightarrow v) + I_f(u) + I_f(w) \leq 2$$

At least one mother:

$$\forall u, v, w : I(\{u, w\} \rightarrow v) - I_f(u) - I_f(w) \leq 0$$

Founder constraints

This always true and provides a speed-up:

$$\forall v : \sum_{v} I(\{\} \rightarrow v) \geq 1$$

Higher lower bounds than 1 can lead to dramatic speed-ups.

Outline

Pedigrees

Integer linear programming

Encoding the pedigree problem

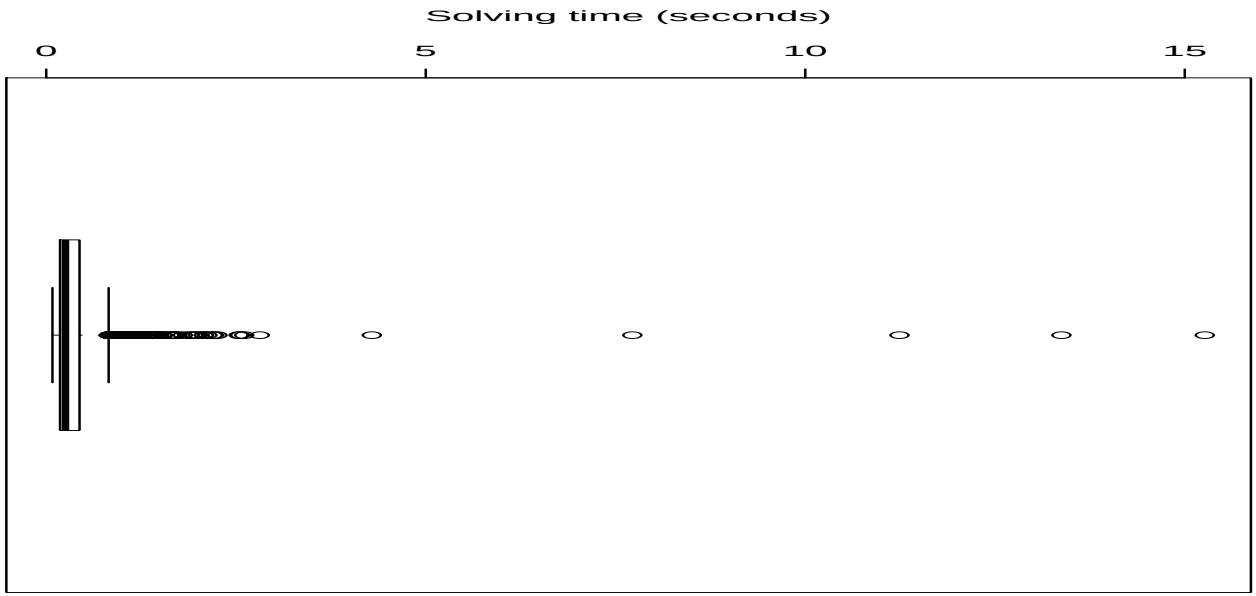
Results

Future directions

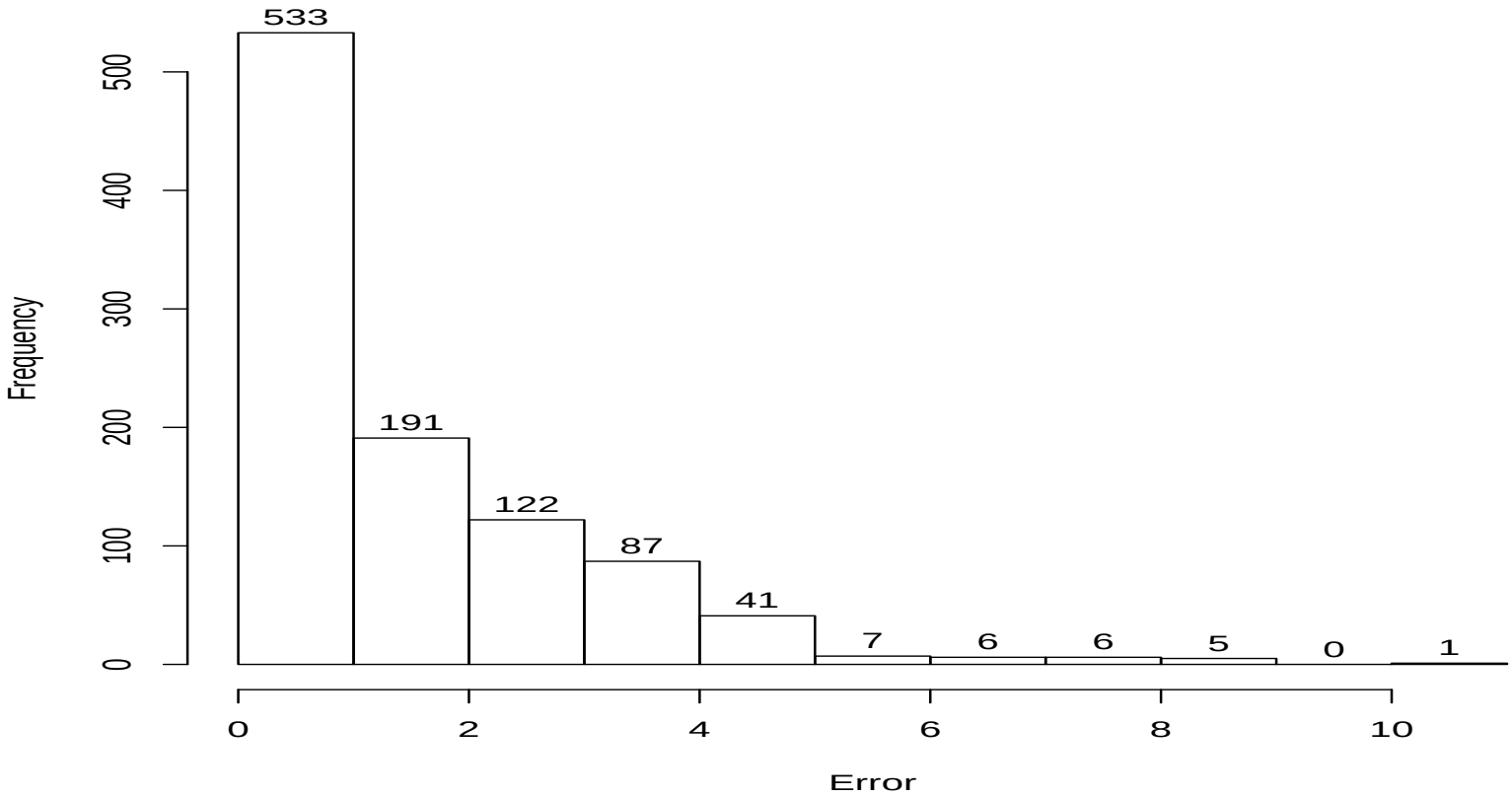
Pedigree reconstruction for 59 individuals

- Datasets were simulated from Almudevar's 59 individual pedigree
- Maximum likelihood pedigrees were obtained from 1000 simulated datasets using the encoding given earlier and the Gurobi ILP solver on a 3GHz dual-core Linux machine.
- The mean solving time was 0.44846 seconds, the median 0.2350 and 75% of runs completed within 0.43860 seconds.
- This is considerably faster than other approaches (and it's exact!)

Distribution of solving times



Distribution of errors



Not always so fast! (46 individual problem)

LB	Time in seconds	Likelihood
0	3189.75858617	-6.3680054000e+02
1	1050.95497108	-6.3680054000e+02
2	1695.93844795	-6.3680054000e+02
3	2350.830338	-6.3680054000e+02
4	1220.98797202	-6.3680054000e+02
5	431.202931166	-6.3680054000e+02
6	246.708929062	-6.3680054000e+02
7	63.2229361534	-6.3680054000e+02
8	3.00327396393	-6.3680054000e+02
9	0.523219823837	-6.3933824000e+02
10	0.293282032013	-6.5144025000e+02

Outline

Pedigrees

Integer linear programming

Encoding the pedigree problem

Results

Future directions

The Bayesian approach

$$\log P(\text{pedigree}|\text{data}) = \log P(\text{pedigree}) + \log P(\text{data}|\text{pedigree}) + K$$

- So for a *maximum a posteriori* (MAP) pedigree need to add extra terms to the objective function.
- This is doable for the log-linear prior of [Sheehan and Egeland, 2007] but requires many auxiliary variables.

Finding the n most probable pedigrees

If G^* is a found maximal probability pedigree, add this constraint

$$\sum_{v \in V} I(\mathrm{Pa}(v, G^*) \rightarrow v) \leq n - 1$$

and start again.

Related work

- [Cowell, 2009] uses a dynamic programming approach on the same problem.
- [Cussens, 2008] uses a weighted MAX-SAT encoding for Bayesian network learning.
- [Jaakkola et al., 2010] use a special-purpose Bayesian network learner which exploits linear relaxations.
- In all cases a limit on the number of parents is required.


Cowell, R. G. (2009).
Efficient maximum likelihood pedigree reconstruction.
Theoretical Popluation Biology, 76(4):285–291.


Cussens, J. (2008).
Bayesian network learning by compiling to weighted MAX-SAT.
In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI 2008)*, pages 105–112, Helsinki. AUAI Press.


Jaakkola, T., Sontag, D., Globerson, A., and Meila, M. (2010).
Learning Bayesian network structure using LP relaxations.
In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*.

Sheehan, N. A. and Egeland, T. (2007).
Structured incorporation of prior information in relationship identification problems.
Annals of Human Genetics.