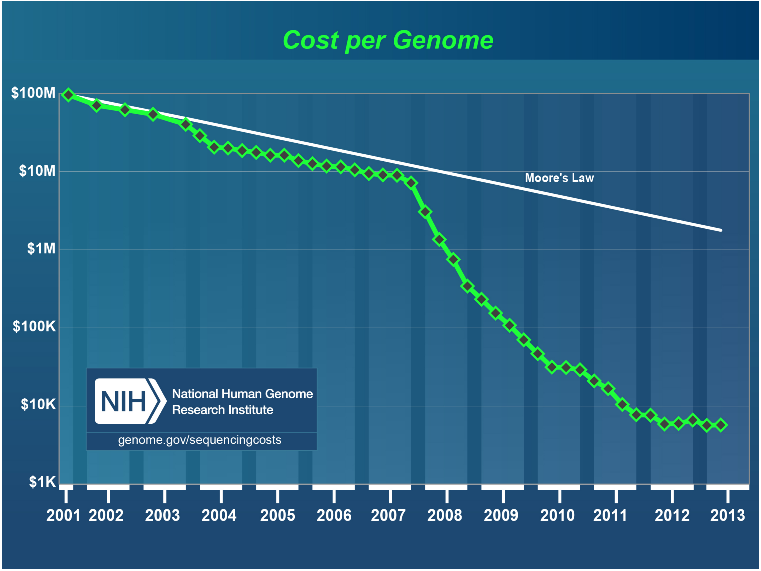
"You're going to need a bigger boat..."

How to stop interesting population genetics models from being swallowed up by really big datasets

Dan Lawson, University of Bristol


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts

Background

- We will ‘soon’ be able to sequence all the genomes in the world for less than the cost of the logistics of obtaining or processing them
- NHS project to sequence 100K people
- Current project to sequence all 50K Faroe Islanders
- What would we do with ‘all the genomes in the world’?
- Can we run appropriate models on them?

Motivation

For Large datasets

- “Statistics doesn’t work” – estimates get worse as we get more data! (for linear compute)
    - e.g. Bayesian models (MCMC)
- Simple analytics can extract many useful features
    - e.g. K-medians clustering, etc
    - Informative in practice - and still hard to get working!
    - But don’t do quite the right thing...
- Many interesting quantities are subtle
- or local, so we only have a small amount of data about them
- Always a place for models closer matching reality

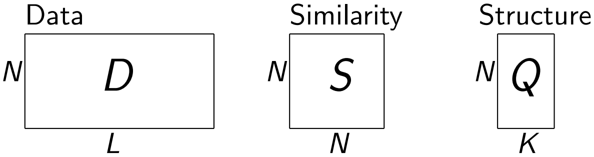
Model of interest: FineSTRUCTURE

Find Populations Q with associated uncertainty from SNP data D

- The N Individuals are highly structured
- The L SNPs are complexly correlated
- Two stage process
- ChomoPainter 'losslessly' describes coancesty $S|D$ using the data
- FineSTRUCTURE infers Q using genetics model $S|Q$
- $S|Q$ is approximately multi-variate normal with structured covariance
- Problem: S is $O(LN^2)$ to evaluate

Similarity

$$p(D|S)p(S|Q)p(Q)$$



- Compare N individuals about which we have lots of genetic data D
- i.e. Painting $S|D$ separates the data D from the population model Q
  - If rows of Q sum to 1 this is a mixture model
  - if only 1 element is non-zero it is a partitioning
- Coancestry $S(i,j)$ is computationally costly to evaluate

Random or convenience filtering

- See 'big data'[1] as better sampling of data
- Why not throw away elements from D?
  - Convenience sampling - what can we measure? $=$'data'
  - Systematic sampling - allele frequency, LD filtering
  - Stratified sampling
  - etc
- For example:
  - Use $L' \ll L$
  - Use $N' \ll N$
- Can fix N' and L' to fix computational cost

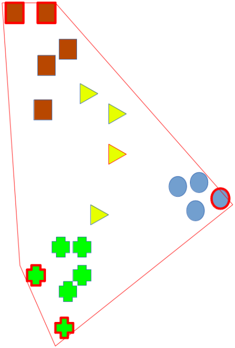
[1] *Big data: any data that can't be processed in memory on a single high spec computer*

Emulated Likelihood Models (ELMs)

- S_{ij} is costly to compute, and needed for $S|Q$
- But are highly structured (e.g. clusters)
- So can emulate (i.e. guess) S_{ij} rather than computing
- Calculate a few S^*, approximately sufficient for $p(D|S,\theta)$
- Carefully downweight emulated values
- Weights are only modification to $S|Q$
- Statistically, emulated values are like Control Variates for the likelihood

Fast finestructure

- Cheap measure S': Use PCA on a few unlinked loci
- Expensive measure S^*: Painting of a few individuals to construct a maximally informative reference panel
- Choose next panel member i_t^* using the *most distant individual to those in the panel*
- Emulation: Predict full paintings $S^\dagger_{\cdot i}$ from panel painting and PCA



How to choose who to paint against whom?

- Iteratively choose the next $S_{.j}$ to add to S^*
- Construct a loss function $\hat{\mathcal{L}}$
- $\hat{\mathcal{L}}$ is implicit here
- Next panel member minimises loss: $\mathrm{argmin}_{S_{ij}} \mathbb{E}\left(\hat{\mathcal{L}}(S^* \cup S_{ij}) | S^*\right)$
- We can consider different histories to evaluate performance
- Stopping rule: convergence of $\hat{\mathcal{L}}$
- $\mathcal{L}$ can represent interest in some populations over others

The emulator as N and L change

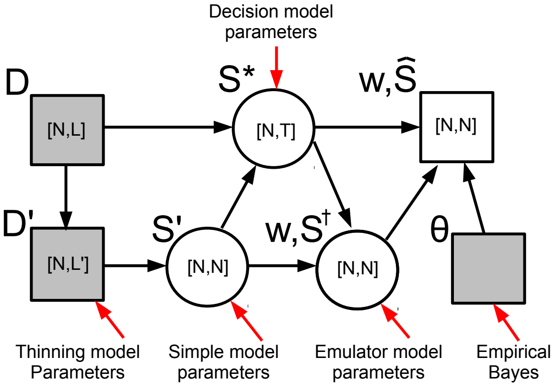
'Machine learning' algorithm with the usual caveats:

- Not a probability model
- Optimal? Unbiased?
- Chosen to respect computational constraints:
  - Lots of loci $L \gg N$: Use PCA for every pair of individuals, and paint a few
  - Lots of individuals $N^2 \gg L$: Computing PCA for every pair of individuals is hard
  - Massive data: Can't even paint everyone against a panel!? There are algorithms that are possible.
- Yet . . . 'Low rank' similarity matrices can be nearly losslessly reconstructed*

*Candes & Plan 'Matrix completion with noise', Proc. IEEE, 2010

Fast finestructure - outline

Emulator:



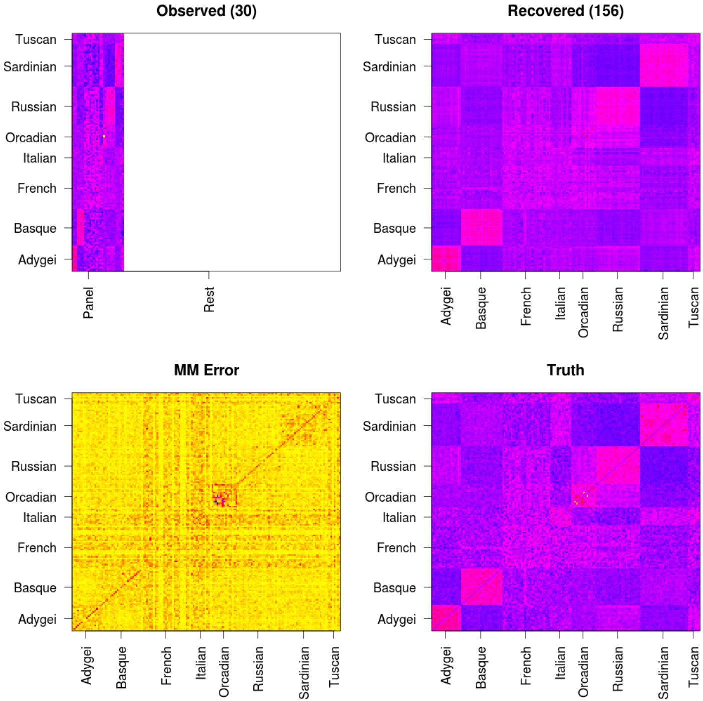
S becomes the data for inferring *Q*:

$$\left[p(D|\hat{S},\hat{\theta})\right] p(S|Q,\phi)p(Q,\phi)$$

Fast finestructure - in practice

- Emulator $\hat{S}$ costs $O(N^2L' + NTL) \ll O(N^2L)$ for full model
- Current datasets: $L = 10,000,000$, $N = 5000$, $L' = 10,000$ $T = 100$, predicted saving ratio is 100
- Bigger savings if we are really only interested in a subset of individuals: can automatically choose an appropriate panel

Thinned PCA approach is cheaper by a factor 10000. L will not grow beyond this, but N will - can introduce an emulation step for S'



Observed (30)

Recovered (156)

MM Error

Truth

Discussion

- Goal: Achieve scale using approximate answers to the right questions via exact answers to the wrong questions
- Proposed the Emulated Likelihood Model:
  - Generally applicable to many problems
  - Full statistical modelling
  - Machine learning algorithms used for the calculation
  - Statistical estimation of parameters is retained but approximated
- FastFineSTRUCTURE is an application (Coming Soon!)
- Take home message for geneticists: You can still develop models that don't scale. Stats is catching up to allow them to scale better.

Thanks for listening!

- Register for FineSTRUCTURE at www.paintmychromosomes.com
- Emulated Likelihood framework developed with Niall Adams, Imperial College London
- FineSTRUCTURE, ChromoPainter work with Garrett Hellenthal, Daniel Falush and Simon Myers

Emulated Likelihood Models for general Bayesian problems

General emulation for big (but not so big) problems

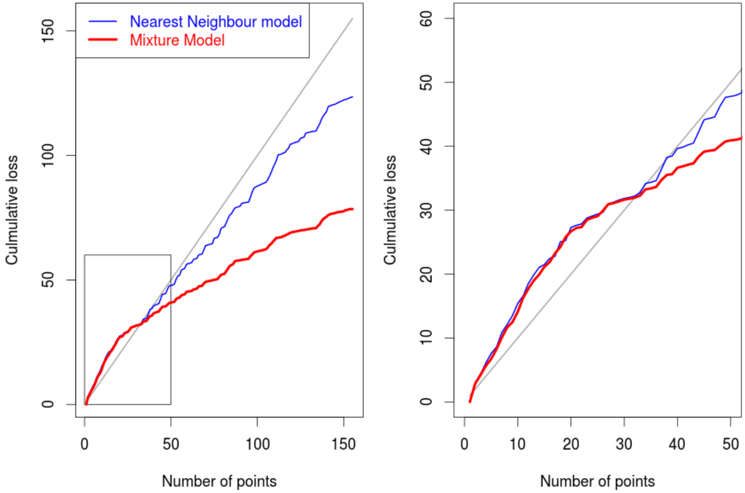
$$\hat{p}(D|S,\theta) = \int p(D|S^* \cup S^\dagger, \theta) p(S^\dagger|S^*, \theta, \psi) d\psi$$

- i.e. Can use θ to emulate $S^\dagger(\theta)$ - e.g. regression in (S, θ) space
  - Gaussian Process for $S_{ij}(\theta)$ is a natural choice
- If $S^\dagger$ is an unbiased estimator of S^* this is a pseudo-marginal approach (and hence targeting the correct posterior)

Fast finestructure - Parallel MCMC algorithm

A parallel tempering algorithm for when MCMC parallelises poorly

- Evaluate the unlinked model S'
- Master node: perform MCMC clustering to find $\hat{Q}_t$ using $\hat{S}_t$, when there are t rows S^*_t computed
- Worker nodes compute $S^*_{\cdot i}$ in the order chosen by the master
- Stopping rule: posterior distribution of $\hat{Q}$ converges
  - No new information added when increasing t
  - (Or if the MCMC is slower than the evaluation of S, sometime afterwards)


Nearest Neighbour model
Mixture Model
Culmulative loss
Number of points
0
50
100
150
0
50
100
150
Culmulative loss
Number of points
0
10
20
30
40
50
60
0
10
20
30
40
50