Estimating Ancestry Coefficients using NMF & Spatial Information

A Comparison to Bayesian Methods

Timo Deist, TIMC-IMAG, Grenoble
Problem Statement

• Population spread over area

• Population structure?

→ Compare genotypes stored in a large matrix $M$

• $M$ encodes absence/presence of specific alleles
  – binary matrix
Pop. Structure Estimation

- Sample Individuals
- Determine genotype
- Estimate number of subpopulations $k$
- Estimate ancestry coefficients
  - similar to clustering
Traditional NMF Objective

\[
\min \left\| M - UV^T \right\|_F^2
\]

s.t. \( \sum_{i \in P_{\ell}} \text{Ancestral Allele Frequencies} \cdot \{ \text{Ancestry Coefficients} \} \)

\[
\sum_{j=1}^{k} V_{ij} = 1 \quad \forall i \in \{1, \ldots, n\}
\]

\( U \geq 0 \)

\( V \geq 0 \)

NMF with Spatial Information

\[
\min \| M - U V^T \|_F^2 + \lambda Tr(V^T L V)
\]

Minimize weighted difference in ancestry coefficients

Spatial Information

\[ w(x, y) = e^{-\frac{d(x, y)}{\sigma}} \]

- \( d(x, y) \): Euclidean distance
- \( \sigma \): determines steepness
  - Set to mean distance of 10 nearest neighbors
Remarks

$$\min \| M - UV^T \|_F^2 + \lambda Tr(V^T LV)$$

- Weighted minimization
  - Difficult to set weights

- Objective is non-convex
  - Global optimum hard to obtain
  - Settle for approximate solutions?
Algorithm

• Alternating Least Squares (ALS)

• Approximate optimal \((U, V)\)
  1) Initialize random \((U, V)\)
  2) Fix one, determine other variable
  3) Vice versa
  4) Repeat until convergence

ALS

$$\min \| M - UV^T \|_F^2 + \lambda Tr(V^T LV) + \beta \sum_{ij} V_{ij}$$

• Observe:
  – convex in either $U$ or $V$
ALS

Repeat until convergence:

1) Fix $V$

2) Solve unconstrained problem for $U$:

$$\min \|M - UV^T\|_F^2 + \lambda \text{Tr}(V^T LV) + \beta \sum_{ij} V_{ij}$$

3) Remap to feasible space

4) Repeat with fixed $U$

5) Repeat until convergence
Data

• Two simulation models:
  – 2-island model with migration & admixture
    • 50 data sets
    • 200 individuals, 1000 SNPs
  – divergence model with admixture
    • 45 data sets
    • 300 individuals, 1000 SNPs

• Real data, Atwell et al. (Science, 2010):
  – Arabidopsis thaliana
    • 170 individuals, 10000 SNPs
Data

• Two simulation models:
  – 2-island model with migration & admixture
    • 50 data sets
    • 200 individuals, 1000 SNPs
  – divergence model with admixture
    • 45 data sets
    • 300 individuals, 1000 SNPs

• Real data, Atwell et al. (Science, 2010):
  – Arabidopsis thaliana
Arabidopsis thaliana

- Model plant with small genome
- Native to Eurasian continent
- Further spread in recent centuries
Arabidopsis thaliana

- 170 European samples with 10k SNPs
- Subset of samples from Atwell et al. (Science, 2010) with 210k SNPs
Results

• Comparison between new ALS algorithm and existing Multiplicative Updates (MU) algorithm
  – MU based on Cai et al. (2010)
Estimating the number of subpopulations $k$

- How many subpopulations $k$ exist?

- Observe estimation error of genotype estimates
  - Root-mean-squared error (RMSE)

- Choose smallest $k$ with significant decrease in RMSE
Estimating the number of subpopulations $k$
Effect of Regularization Parameters on Ancestry Estimation

• **ALS**: Strictly better estimates when including geographic information
Estimation Error of Ancestry Coefficients

• Ancestry coefficients estimated in $V$

• Compare RMSE of estimates vs. true values
  – True values are known for simulated data

• Comparison of best parametrizations
Estimation Error
Results – MU versus ALS

• RMSE is lower (ALS)

• Relatively stable optimal parameterization (ALS): \((\beta, \lambda) = (50, 100)\)

• Similar convergence time
Results

• Comparison with TESS, Durand et al. (MBE, 2009)
  – estimation error (sim. Data)
  – ancestry coefficients for *Arabidopsis thaliana*

• TESS
  – models distribution of ancestry coefficients
  – estimates based on Monte Carlo Markov Chains
Estimation Error

Tess (red) & ALS (black)

RMSE

Divergence Time Subpop. II & III $\epsilon_{23}$
Arabidopsis thaliana

François et al. (2008), Demographic History of European Populations of Arabidopsis thaliana. PLOS Genetics
Results – ALS versus TESS

ALS

• Consistently lower average RMSE (sim. data)

• Lower execution time (real data)
  – ALS: < 6 seconds
  – TESS: 170 seconds
Conclusion

• Geographic information improves estimates
  – Parameterization is difficult

• NMF is fit for large data sets
  – Significantly faster than TESS

• Non-parametric approach with same interpretation as model-based approach