FLPCA: A Fused Lasso PCA-based approach to identify influential markers in differentiated populations from dense SNP data
Denis Laloë, Julien Chiquet, Florence Jaffrézic, Mathieu Gautier

To cite this version:

HAL Id: hal-01075342
https://hal.archives-ouvertes.fr/hal-01075342
Submitted on 17 Oct 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
FLPCA: A Fused Lasso PCA-based approach to identify influential markers in differentiated populations from dense SNP data

IBC meeting
Firenze, July 2014

Denis Laloë, Julien Chiquet, Florence Jaffrézic, Mathieu Gautier

July 8, 2014
The context

Genetic structure of a population
- Natural / Artificial Selection
- Isolation, drift

Markers: SNPs
- Usually biallelic markers
- Throughout the genome
- Mapping SNPs to genes

Geometric Data Analysis
- Duality Diagram: Space of individuals vs Space of markers
- Typological Value
- Modelling through instrumental variables

Subset of influential markers
- Magnitude
- Spatial structure (Linkage Disequilibrium)
SNP Haplotype data

\[ X = [\delta_i^j] = \begin{pmatrix}
SNP_1 & SNP_2 & \ldots & SNP_k \\
0 & 1 & \ldots & 0 \\
0 & 0 & \ldots & 1 \\
\vdots & \vdots & \ddots & \vdots \\
1 & 0 & \ldots & 0
\end{pmatrix} \]
Duality diagram

Dray and Dufour, 2007

Maximisation of the correlation between variables and components

\[ V = X'X/n \]

Variables

Maximisation of the individuals dispersion

\[ W = XX'/n \]

Individus

Diagonalisation

\[ X'X = AA \]
\[ A'A = I \]

Principal components

same non-null eigenvalues

\[ \lambda_1 > \lambda_2 > \ldots > \lambda_r > 0 \]
\[ \Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r) \]

Transition formulae

\[ XA^{0.5} = B \]
\[ X'B^{0.5} = A \]

Coordinates of variables

C = X'B

Coordinates of individuals

L = XA

Principal axes
Typological Values

- **Haplotypes**
  - coordinates of haplotypes along the ith axis $y_i$

- **SNPs**
  - coordinates of SNPs along the ith component $c_i$
  - $c_{ij}^2$: Typological value of $SNP_j$ with the ith component
    - $R^2$ of the model
      $y_i = \mu + SNP_j + \epsilon$
    - $Fst$ (Laloë and Gautier, 2011)
Problem formulation : the Fused Lasso Signal Approximator (FLSA)

(Tibshirani et al, 2005; Hoefling, 2010)

- $\mathbf{y} = (y_1, \ldots, y_n)$ an ordered vector of data
- identification of consecutive points with high and constant values.
- FLSA solution

$$\hat{\beta}(\lambda_1, \lambda_2) = \arg\min_{\beta} \left\{ \frac{1}{2} \| \mathbf{y} - \beta \|^2 + \lambda_1 \sum_{i=1}^{n} |\beta_i| + \lambda_2 \sum_{i=1}^{n-1} |\beta_{i+1} - \beta_i| \right\}$$

- $\lambda_1$ controls the level of sparsity
- $\lambda_2$ controls the level of smoothness
Adaptive FLSA

A two step procedure (Rinaldi, 2009).

1. Fusion step
   1.1 Fit the FLSA model with $\lambda_1 = 0$, i.e., $\hat{\beta}(0, \lambda_2)$.
   1.2 For the partition $\mathcal{B} = \{B_1, \ldots, B_J\}$ of $J$ blocks (or segments) associated to $\hat{\beta}(0, \lambda_2)$, compute:

   $\tilde{\beta} = \sum_{j=1}^{J} \bar{y}_j 1_{B_j}, \quad \bar{y}_j = \text{card}(B_j)^{-1} \sum_{j=1}^{J} y_j,$

2. Adaptive step
   Fit the following weighted lasso problem:

   $\hat{\beta}^{AFL} = \arg \min_{\beta} \left\{ \frac{1}{2} \left\| \beta - \tilde{\beta} \right\|^2_2 + \lambda_1 \sum_{i=1}^{n} w_i |\beta_i| \right\}, \quad w_i = \sum_{j=1}^{J} \frac{1_{B_j}}{\sqrt{\text{card}(B_j)}}$

Hard thresholding : hard$(x; \lambda) = x \cdot 1_{\{|x| > \lambda\}}$
Model selection

- Cross Validation: Construction of fold not obvious in the case of ordered data
- Penalized Criterion
Penalized Criteria

- IC penalized criterion of the form

\[
\text{IC}(\hat{\beta}(\lambda_1, \lambda_2)) = \frac{1}{2n} \left\| y - \hat{\beta} \right\|^2 + \frac{\sigma^2}{n} \text{pen}(\text{df}),
\]

where

- \text{pen}(\cdot) is a function that penalized the number of parameters of the model, df (number of segments different from zero).
- \(\sigma^2\) estimated by the plug-in estimator of P Hall (Lebarbier, 2005)

- Criteria

- AIC: \(\text{pen}(\text{df}) = 2 \cdot \text{df}\)
- BIC: \(\text{pen}(\text{df}) = \log(n) \cdot \text{df}\)
- BML (for Birgé, Massart, Lebarbier; Lebarbier, 2005) : \(\text{pen}(\text{df}) = \text{df} \ast (2 \ast \log(n/\text{df}) + 5)\)
Simulations

- Three populations
- Five 5 Mb-chromosomes
- 1000 SNPs per chromosome
- 1 causal variant at position 2.5 Mb of chromosome 1
  driven to fixation via selection in population P1
- Program msms coalescent simulator *Ewing and Hermison, 2010*
- 100 simulations
- First PC: Highest node
  P1-P2 vs P3
A simulation example. Parameter optimization

Choice of $\lambda_1$ and $\lambda_2$ according to the penalization criterion
A simulation example. Selected regions

Selected SNPs according to the penalization criterion
Comparison of penalization criteria

1. Choice of parameters

2. Selected regions
French bovine data

- 600 animals from 20 French dairy and beef cattle breeds
- HD 770k SNP array
- PCA on haplotypes
- a between **beef** vs **dairy** analysis
Results

- 10675 selected markers (out of 656152): 1.6%
- Accounts for spatial structure, selected regions containing from 1 to 76 SNPs (\( \mu = 2.5 \))

![Magnitude of selected SNPs according to the block length](image_url)
Results

Manhattan plot of typological values. **Selected SNPs**
Results

- 1067 genes containing selected markers
- Enrichment analysis on the first 100 genes
  - Cytoskeletal protein binding (Muscle)
  - Lipid metabolism process (Milk/Muscle)
Conclusions and Future Work

- Selection of 1.6% markers
- Accounts for spatial structure, selected regions containing from 1 to 76 SNPs ($\mu = 2.5$)
- Sensitivity to penalization criterion
- Sensitivity to parameter tuning

- Stability selection strategy (Meinshausen and Bühlmann, 2010; Yang et al, 2011)
  - Reducing false positives
  - Reducing the effect of parameter tuning
Acknowledgements

Bovine data provided by

- ANR project GEMBAL

Funded by

- ANR project EpiGrani
- Métaprogramme INRA-ACCAF project GALIMED
References