SoCal: supervised genotype calling via ellipsoidal separation for Affymetrix SNP microarray

Huwenbo Shi (603-778-363) shihuwenbo@ucla.edu

In this article, I present SoCal, a supervised genotype calling algorithm for Affymetrix SNP microarray. For each SNP, SoCal first efficiently identify ellipsoidal decision regions for each genotype from reference genotype calls, and then uses these regions to classify future SNPs into different genotypes. Using only a small portion of training genotype calls from the HapMap Project, SoCal achieves an accuracy of 97.5% during validation.

1 Introduction

Accurate genotyping of SNPs is essential to discovering true signals in association studies. Although next generation sequencing technology provides cheap whole–genome sequences for genotyping SNPs, SNP microarray is still a cost–effective genotyping technology for many specific association studies. In an Affymetrix SNP microarray, oligonucleotide probes are used to match and bind DNA fragments containing biallelic SNPs. Then a fluorescence scanner scans the microarray to quantify perfect match and mismatch of these fragments. Most genotype calling procedures for SNP microarray consists of two steps. In the first step, raw information from microarray is summarized to obtain the intensities, θ_A and θ_B, of the two alleles, denoted by A and B, of each SNP. In the second step, SNPs are classified into genotype AA, AB, or BB based on the allele intensities they generate. The focus of this article is on the second step of the genotype calling procedure—genotype calling using summarized allele intensities.

For a specific SNP, if a sample has genotype AA or BB, the allele intensity, θ_A or θ_B, will be higher respectively. If a sample has genotype AB, the intensities, θ_A and θ_B, will be similar. If one plots $log(\theta_A)$ versus $log(\theta_B)$ of a SNP for a number of samples, normally 3 ellipsoidal clusters can be observed, one for each genotype, as shown in Figure 1. Many genotype calling algorithms use model–based unsupervised clustering to identify these clusters and then assign genotypes to each cluster. Although these methods are applicable to a wide range of microarrays, they don't take advantage of genotype calls that are already available. Also, these methods use the EM algorithm to estimate model parameters, which is sensitive to starting parameters and slow to converge. To utilize reference genotype calls, Rabbee et al. proposed the RLMM algorithm, a supervised genotype calling method that uses reference genotype calls to form decision regions by fitting bivariate Gaussian distributions on observed allele intensities for each genotype.

These Gaussian decision regions are then used to call SNPs with unknown genotype. However, fitting a Gaussian distribution is known to be sensitive to outliers, which, for SNP microarrays, can be caused by genomic structural variations.

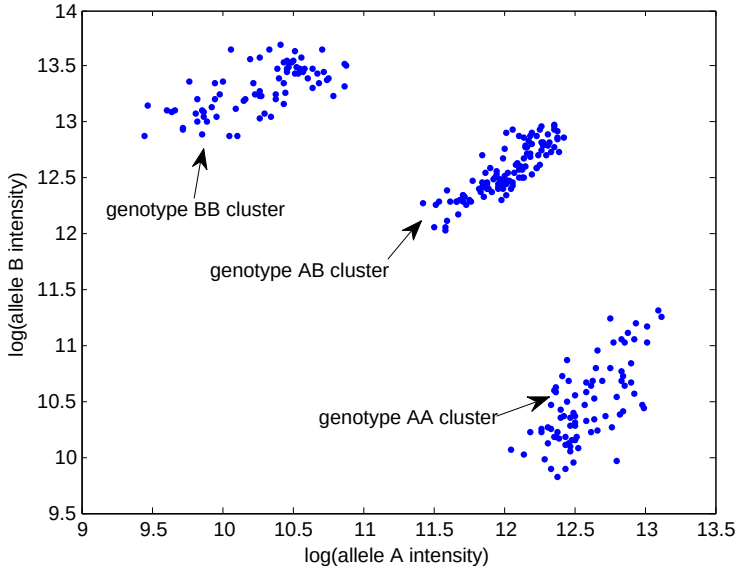


Figure 1: Genotype clusters obtained from Affymetrix SNP array allele intensity values

In this article, I present SoCal, a supervised genotype calling algorithm for Affymetrix SNP microarrays. Instead of fitting Gaussian distributions on allele intensities with reference genotype calls, SoCal efficiently finds ellipsoidal decision regions for each genotype of a SNP by solving a conic programming problem. SoCal can control the effect of outliers by assigning different weights to each of the criteria for finding the ellipsoids—separation ratio, ellipsoid volume, and number of enclosed points. After SoCal finds the ellipsoidal decision regions for each genotype of a SNP, it uses them to call the same SNP with unknown genotype.

Using reference genotype calls from the HapMap Project as training and validation data, SoCal achieved 99.71% accuracy at a call rate of 95% during leave-one-out cross-validation. Furthermore, SoCal shows more robustness than the RLMM method when there are outliers in the training data.

2 Method

2.1 Overview of SoCal's genotype calling procedure

SNP allele intensities are first summarized from raw microarray data using SNPRMA, which removes non–biological effect from the data. After this step, SoCal calls genotypes in two steps. In the first step, SoCal finds ellipsoidal decision regions for each of the

genotype of a SNP using reference genotype calls. In the second step, SoCal classifies SNPs with unknown genotypes using minimum distance classification.

2.2 Pattern separation by ellipsoid

An ellipsoid $\mathcal{E} \subseteq \mathbb{R}^n$ can be expressed as $\mathcal{E} = \{x \in \mathbb{R}^n | (x-c)^T E(x-c) \leqslant 1\}$, where c is the center of the ellipsoid, and E a positive definite matrix denoting the shape and orientation of the ellipsoid. Let $\{a_i\}$ be the points to be included in an ellipsoid, and $\{b_j\}$ be the points to be excluded, the problem of ellipsoidal separation is to find c and E such that $(a_i - c)^T E(a_i - c) \leqslant 1 \, \forall i$ and $(b_j - c)^T E(b_j - c) > 1 \, \forall j$.

2.3 Forming ellipsoidal decision regions for each genotype

Let $G = \{AA, AB, BB\}$ be the set of genotypes of a SNP, and J_{AA}, J_{AB}, J_{BB} the index set of samples with the corresponding genotype. Let $X = \{(log(\theta_A), log(\theta_B))_i | i = 1, \cdots, |J_{AA}| + |J_{AB}| + |J_{BB}|\}$ be the set of log transformed allele intensities of all the samples, and $X_{AA} = \{x_j | x_j \in X, j \in J_{AA}\}$, $X_{AB} = \{x_j | x_j \in X, j \in J_{AB}\}$, $X_{BB} = \{x_j | x_j \in X, j \in J_{BB}\}$ the set of log transformed allele intensities from samples having the corresponding genotype.

To find the ellipsoid that includes X_{AA} and excludes $X_{AB} \cup X_{BB}$, one sets $\{a_i\} = X_{AA}$ and $\{b_j\} = X_{AB} \cup X_{BB}$, and solves the following conic programming problem. For the sake of space, detailed derivation of the problem formulation is not presented here.

$$
\begin{aligned}
\text{minimize} \quad & -\beta_1 k + \beta_2 trace(T) + \beta_3 \|u - \mathbb{1}\|_1 \\
\text{subject to} \quad & (1, a_i)^T \tilde{E} (1, a_i) \leqslant u_i \, \forall i \\
& (1, b_j)^T \tilde{E} (1, b_j) \geqslant k \, \forall j \\
& \tilde{E} = \begin{bmatrix} s & v^T \\ v & F \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} F & I \\ I & T \end{bmatrix} \succeq 0
\end{aligned}
$$

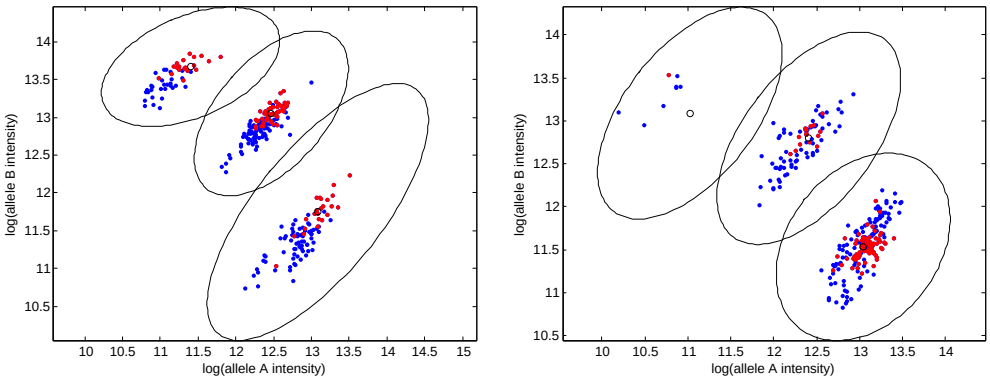
In the problem formulation above $\beta_i > 0$ are the weights assigned to each criteria of finding the ellipsoid—separation ratio, ellipsoid volume, and number of enclosed points. In SoCal, β_1, β_2, β_3 are empirically set to 1, 10^4, and 10^2 respectively.

Let $\tilde{E}^* = \begin{bmatrix} s & v^T \\ v & F \end{bmatrix}$ be the optimal solution to the problem above. The separating ellipsoid $\mathcal{E}^*$ is defined as $\mathcal{E}^* = \{x \in \mathbb{R}^n | (x - c^*)^T E^* (x - c^*) \leqslant 2(1 + k)\}$, where $c^* = -F^{-1}v$, $E^* = \frac{F}{(1 - s + c^{*T} F c^*)}$.

To find the ellipsoid that includes X_{AB} and excludes $X_{AA} \cup X_{BB}$, one sets $\{a_i\} = X_{AB}$ and $\{b_j\} = X_{AA} \cup X_{BB}$, and solves the above conic programming problem. The same procedure also applies to finding the ellipsoid that includes X_{BB} and excludes $X_{AA} \cup X_{AB}$.

2.4 Rescuing missing genotype clusters

If a SNP has moderate minor allele frequency (MAF), the genotype clusters of that SNP are well defined, and SoCal obtains three ellipsoidal decision regions for that SNP, one for each genotype (Figure 2a). However, if a SNP has lower MAF, some genotype cluster may not be well defined. For these SNPs, SoCal estimates the ellipsoid for the missing genotype cluster using the ellipsoids for the other two genotypes through simple geometric transformations (Figure 2b).



(a) SNP with well defined genotype clusters (b) SNP with sparse BB genotype cluster

Figure 2: Each dot in the plots above represents a sample, with samples having HapMap reference genotype calls marked red. The ellipsoids were obtained using only the reference calls.

2.4.1 Missing genotype AA or BB cluster

If genotype AA cluster of a SNP has less than 3 reference calls, SoCal first finds the ellipsoids for genotype AB and BB clusters, and then estimates that for genotype AA cluster through simple geometric transformations.

Let $\mathcal{E}_{AB} = \{x \in \mathbb{R}^n | (x - c_{AB})^T E_{AB} (x - c_{AB}) \leqslant 1\}$ and $\mathcal{E}_{BB} = \{x \in \mathbb{R}^n | (x - c_{BB})^T E_{BB} (x - c_{BB}) \leqslant 1\}$ be the ellipsoids obtained for genotype AB and BB clusters, and n_{AB}, n_{BB} the unit vectors pointing in the direction of the major axis of the corresponding ellipsoid. SoCal estimates the center of $\mathcal{E}_{AA}$, the ellipsoid for genotype AA cluster, by reflecting c_{BB}, the center of $\mathcal{E}_{BB}$, across the major axis of $\mathcal{E}_{AB}$. To estimate the orientation of $\mathcal{E}_{AA}$, SoCal first determines the angle between n_{AB} and n_{BB}, and then applies a rotation matrix of that angle on E_{AB}.

Formally, let $\mathcal{E}_{AA} = \{x \in \mathbb{R}^n | (x - c_{AA})^T E_{AA} (x - c_{AA}) \leqslant 1\}$ be the estimated ellipsoid for genotype AA cluster, and α the angle between n_{AB} and n_{BB}, then $c_{AA} = -c_{BB} + 2c_{AB} + 2n_{AB}((c_{BB} - c_{AB})^T n_{AB})$, and $E_{AA} = R^T E_{AB} R$, where R is a rotation matrix of angle α.

If genotype BB cluster is missing, the center and orientation of the ellipsoid for that cluster is estimated in a similar way. Formally, let $\mathcal{E}_{BB} = \{x \in \mathbb{R}^n | (x - c_{BB})^T E_{BB} (x -$

$c_{BB}) \leqslant 1\}$ be the estimated ellipsoid for genotype BB cluster, and α the angle between n_{AB} and n_{AA}, then $c_{BB} = -c_{AA}+2c_{AB}+2n_{AB}((c_{AA}-c_{AB})^T n_{AB})$, and $E_{BB} = R^T E_{AB} R$, where R is a rotation matrix of angle $-\alpha$.

2.4.2 Missing genotype AB cluster

Although SNPs with genotype AB cluster missing were not observed in HapMap reference genotype calls, for completeness, for these SNPs SoCal first obtains, $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$, the ellipsoids for genotype AA and BB cluster, and then estimates the center of $\mathcal{E}_{AB}$, the ellipsoid for the missing cluster, using the mid-point between the centers of $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$. The orientation of $\mathcal{E}_{AB}$ is obtained by applying a rotation to the ellipsoid with the minimum volume among $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$.

Formally, let $\mathcal{E}_{AB} = \{x \in \mathbb{R}^n | (x - c_{AB})^T E_{AB} (x - c_{AB}) \leqslant 1\}$ be the estimated ellipsoid for genotype AB cluster, and α the angle between n_{AA} and n_{BB}, then $c_{AB} = (c_{AA} + c_{BB})/2$, and $E_{AB} = R^T \hat{E} R$, where $\hat{E}$ is the matrix of the ellipsoid with the minimum volumne among $\mathcal{E}_{AA}$ and $\mathcal{E}_{BB}$, and R a rotation matrix of angle $\pm\alpha/2$. The sign of the angle of rotation is dependent on the choise of ellipsoid on which rotation is applied—positive for $\mathcal{E}_{AA}$ and negative for $\mathcal{E}_{BB}$.

2.5 Genotype calling

After the ellipsoidal decision regions, $\mathcal{E}_g = \{x \in \mathbb{R}^n | (x - c_g)^T E_g (x - c_g) \leqslant 1\}, \forall g \in \{AA, AB, BB\}$ of a SNP are obtained, SoCal uses them to classify SNPs with unknown genotypes using minimum distance classification.

If a sample has allele intensity θ_A and θ_B at SNP n, SoCal first computes $D_g = \sqrt{(x - c_g)^T E_g (x - c_g)}$, where $x = (log(\theta_A), log(\theta_B))$, for each $g \in \{AA, AB, BB\}$. SoCal then calls the genotype, $\mathcal{G}$, of that sample at SNP n as the genotype having the minimum D_g, that is, $\mathcal{G} = \arg\min_{g \in \{AA,AB,BB\}} D_g$. SoCal defines $\lambda = 1 - D_{\mathcal{G}}/(D_{AA} + D_{AB} + D_{BB})$ to quantify the confidence of each genotype call.

3 Data

The microarray used for evaluation in this project was the Affymetrix GeneChip Human Mapping 50K Xba Array, which contains 58,960 SNPs. Raw microarray data for 270 samples was obtained from HapMap FTP. And reference genotype calls were obtained using HapMart.

After removing strand-ambiguous SNPs and SNPs not present on HapMart from the microarry, 16,387 SNPs were left. Figure 3 shows the minor allele frequency distribution for these SNPs.

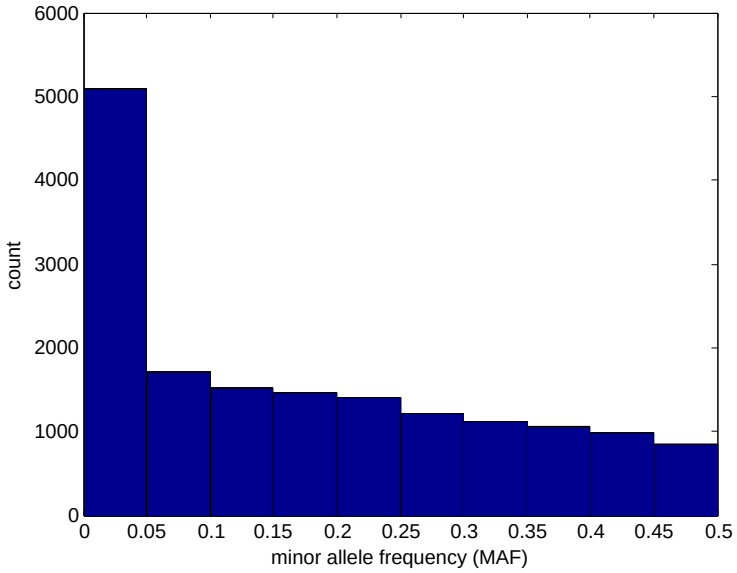


Figure 3: Minor allele frequency distribution for the 16,387 SNPs.

4,064 SNPs with two genotype clusters having less than 3 reference genotype calls were further removed from the microarry. Among these SNPs, 3,596 are monomorphic SNPs. In total, 12,323 SNPs were left for evaluation.

4 Result

4.1 Comparison with HapMap reference calls

HapMap/SoCal	AA	AB	BB	No Call
AA	360,289	2,282	1,058	0
AB	2,667	341,012	2,257	0
BB	851	2,347	368,556	0

Table 1: At a call rate of 100%, SoCal achieved 98.94% concordance rate in the leave-one-out cross-validation with HapMap reference calls.

HapMap/SoCal	AA	AB	BB	No Call
AA	348,221	390	298	14,720
AB	710	319,394	775	25,057
BB	410	427	357,627	13,290

Table 2: At a call rate of 95%, SoCal achieved 99.71% concordance rate in the leave-one-out cross-validation with HapMap reference calls.

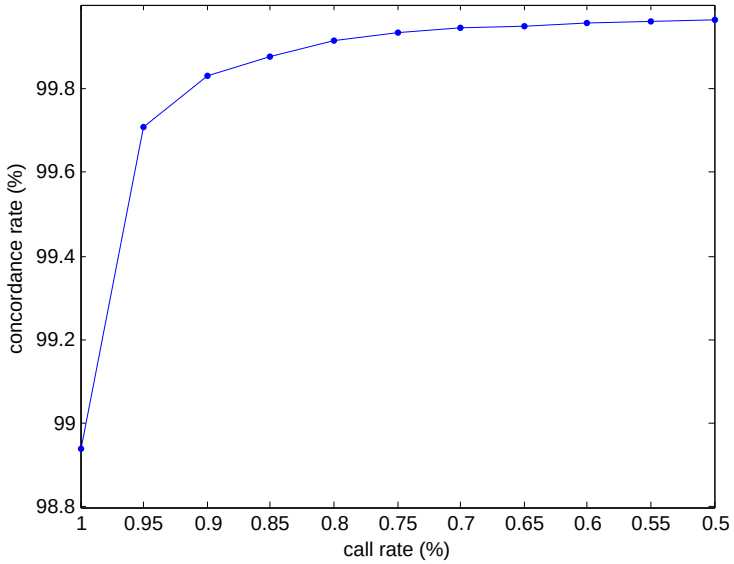
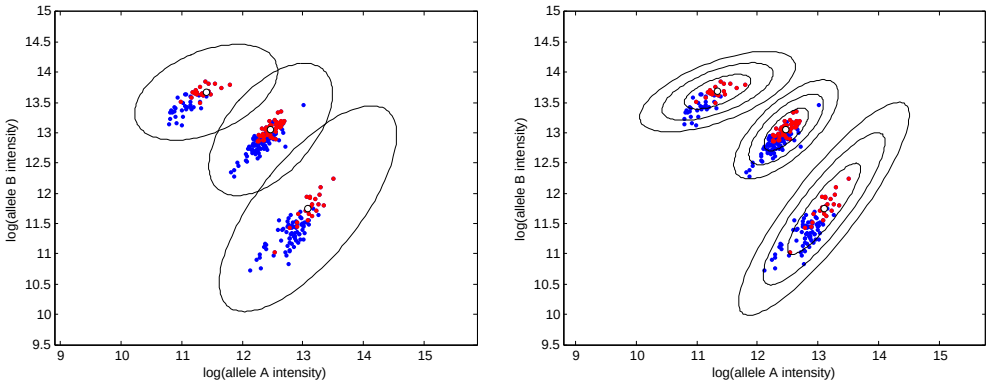


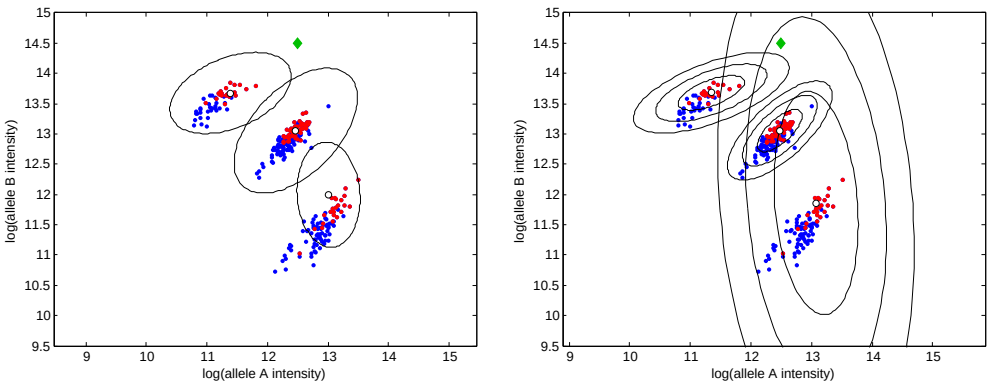
Figure 4: Concordance rate of SoCal in the leave-one-out cross-validation with HapMap reference calls as a function of call rate.

4.2 Comparison with RLMM



(a) Decision regions found by SoCal when there is no outlier

(b) Decision regions found by RLMM when there is no outlier



(c) Decision regions found by SoCal when there is no outlier

(d) Decision regions found by RLMM when there is no outlier

Figure 5: Each dot in the plots above represents a sample, with samples having HapMap reference genotype calls marked red. The ellipsoids were obtained using only the reference calls.

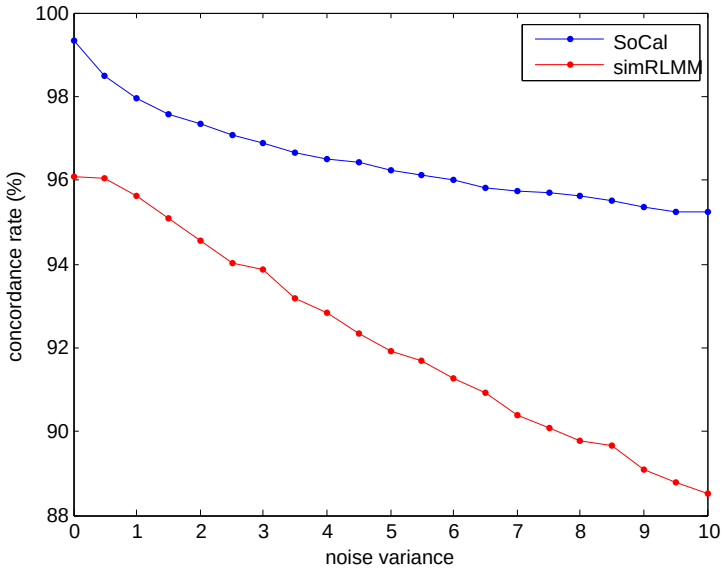


Figure 6: Concordance rate of SoCal and RLMM in the leave-one-out cross-validation with HapMap reference calls as a function of noise variance.

4.3 Comparison with CRLMM calls

TODO: Compare with CRLMM concordance rate call rate

5 Discussion

TODO: Write discussion section

References

[1] Rho, S. W., Abell, G. C., Kim, K., Nam, Y., & Bae, J. (2010). Comparing microarrays and next-generation sequencing technologies for microbial ecology research. Trends in Biotechnology, 28, 291-299.

[2] Norlén, H., Pettersson, E., Ahmadian, A., Lundeberg, J., & Sundberg, R. (2008). Classification of SNP genotypes by a Gaussian mixture model in competitive enzymatic assays. Mathematical Statistics Stockholm University Research Report, 3, 1-26.

[3] Lin, Y., Tseng, G. C., Cheong, S. Y., Bean, L. J., Sherman, S. L., & Feingold, E. (2008). Smarter clustering methods for SNP genotype calling. Bioinformatics, 24, 2665-2671.

[4] Wu, C. F. (1983). On the Convergence Properties of the EM Algorithm. The Annals of Statistics, 11, 95-103.

[5] Rabbee, N., & Speed, T. P. (2005). A genotype calling algorithm for affymetrix SNP arrays. Bioinformatics, 22, 7-12.
[6] Zhou, X., & Stephens, M. (2014). Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nature Methods, 11, 407-409.
[7] Glineur F. (1998). Pattern separation via ellipsoids and conic programming. (MS Thesis). Facult Polytechnique de Mons, Mons, Belgium.